New York City's
Property Valuation Analysis

Table of Contents

I. Executive Summary

This project investigates property tax fraud in New York City by utilizing advanced fraud analytics on a dataset of approximately 1 million property records provided by the Department of Finance. The dataset includes detailed information on property valuations, characteristics, and ownership, with the objective of identifying anomalies that could indicate potential fraudulent activity.

The analysis followed a systematic approach, beginning with data cleaning to address missing values, retain outliers as potential anomalies, and filter for private properties relevant to the investigation. New variables were engineered to improve anomaly detection, including valuation ratios, inverse metrics, and regional averages. Principal Component Analysis (PCA) was applied for dimensionality reduction, ensuring the data's complexity was manageable while preserving key information. Two unsupervised anomaly detection algorithms—Z-Score Outliers and Autoencoder—were employed to generate fraud scores by identifying deviations in property metrics.

The results uncovered significant anomalies, with the most unusual records integrated back into the dataset for further examination. Visualizations, including bar charts and heatmaps, highlighted key anomalies and the most impactful variables. This data-driven approach demonstrates the potential to enhance fraud detection systems and protect public revenue by identifying and addressing discrepancies in property tax assessments.

II. Dataset Description

1. Dataset Description

This dataset contains New York City's Property Valuation and Assessment Data provided by the Department of Finance (DOF). It includes approximately 1 million property records across 32 fields, combining categorical and numerical variables, with some missing values. Each record represents a unique property in the city's five boroughs, providing information such as address, owner details, property characteristics, and assessed tax values.

2. Field Summary Tables

1) Numeric fields

Field Name	Field Type	# Records Have Values	% Populated	# Zeros	Min	Max	Mean	Standard Deviation	Most Common
LTFRONT	numeric	1,070,994	100.0%	169,108	0	9999	37	74	0
LTDEPTH	numeric	1,070,994	100.0%	170,128	0	9999	89	76	100
STORIES	numeric	1,014,730	94.7%	0	1	119	5	8	2
FULLVAL	numeric	1,070,994	100.0%	13,007	0	6,150,000,000	874,265	11,582,426	0
AVLAND	numeric	1,070,994	100.0%	13,009	0	2,668,500,000	85,068	4,057,258	0
AVTOT	numeric	1,070,994	100.0%	13,007	0	4,668,308,947	227,238	6,877,526	0
EXLAND	numeric	1,070,994	100.0%	491,699	0	2,668,500,000	36,424	3,981,574	0
EXTOT	numeric	1,070,994	100.0%	432,572	0	4,668,308,947	91,187	6,508,400	0
BLDFRONT	numeric	1,070,994	100.0%	228,815	0	7,575	23	36	0
BLDDEPTH	numeric	1,070,994	100.0%	228,853	0	9,393	40	43	0
AVLAND2	numeric	282,726	26.4%	0	3	2,371,005,000	246,236	6,178,952	2,408
AVTOT2	numeric	282,732	26.4%	0	3	4,501,180,002	713,911	11,652,508	750
EXLAND2	numeric	87,449	8.2%	0	1	2,371,005,000	351,236	10,802,151	2,090
EXTOT2	numeric	130,828	12.2%	0	7	4,501,180,002	656,768	16,072,449	2,090

1.1) Property Dimensions (Lot and Building Sizes)

- Fields: LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH
- Description: These fields represent the dimensions of properties, such as lot width, lot depth, building width, and building depth. Values range widely (0 to thousands), with averages around 20-90 feet. All four fields have significant numbers of records with zeros (ranging from 169,108 to 228,853).

1.2) Property Valuations (Market, Actual, and Excess Values)

- Fields: FULLVAL, AVLAND, AVTOT, EXLAND, EXTOT
- Description: These fields represent different valuation metrics, such as market value (FULLVAL), actual land and total values (AVLAND, AVTOT), and actual exempt land and total values (EXLAND, EXTOT). Values range from $0 to billions, with high averages (e.g., $874,265 for FULLVAL) and significant variability (standard deviations in millions). Many records contain zeros, which may indicate properties with no assessed or excess value.

1.3) Secondary Valuations (Transitional Assessments)

- Fields: AVLAND2, AVTOT2, EXLAND2, EXTOT2
- Description: These fields represent transitional valuation metrics, associated with specific adjustments or exemptions applied to a subset of properties. They reflect tax policies or temporary states in property valuation. Values range from $0 to billions, similar to the primary valuation fields, but they have significantly lower population rates in comparison (e.g., 26.4% for AVLAND2 and 8.2% for EXLAND2)

1.4) Building Characteristics

- Fields: STORIES
- Description: This field represents the height of a building, measured by the number of stories. Values range from 1 to 119, with an average of 5 stories. It is 94.7% populated, making it a generally reliable indicator of building structure.

2) Categorical fields

Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
BBLE	categorical	1,070,994	100.0%	0	1,070,994	1000010101
BORO	categorical	1,070,994	100.0%	0	5	4
BLOCK	categorical	1,070,994	100.0%	0	13,984	3944
LOT	categorical	1,070,994	100.0%	0	6,366	1
EASEMENT	categorical	4,636	0.4%	0	12	E
OWNER	categorical	1,039,249	97.0%	0	863,347	PARKCHESTER PRESERVAT
BLDGCL	categorical	1,070,994	100.0%	0	200	R4
TAXCLASS	categorical	1,070,994	100.0%	0	11	1
EXT	categorical	354,305	33.1%	0	3	G
EXCD1	categorical	638,488	59.6%	0	129	1017
STADDR	categorical	1,070,318	99.9%	0	839,280	501 SURF AVENUE
ZIP	categorical	1,041,104	97.2%	0	196	10314
EXMPTCL	categorical	15,579	1.5%	0	14	X1
EXCD2	categorical	92,948	8.7%	0	60	1017
PERIOD	categorical	1,070,994	100.0%	0	1	FINAL
YEAR	categorical	1,070,994	100.0%	0	1	2010/11
VALTYPE	categorical	1,070,994	100.0%	0	1	AC-TR

2.1) Unique Identifiers

- Fields: BBLE, BLOCK, LOT
- Description: These fields act as unique identifiers for properties. BBLE represents the file key, which is a combination of borough, block, lot, and easement codes. It is completely unique across all records, with 1,070,994 unique values. BLOCK indicates the block number within five boroughs. It has 13,984 unique values and is fully populated. LOT represents the lot number within a block, with 6,366 unique values. This field is also 100% populated.

2.2) Location and Address

- Fields: BORO, STADDR, ZIP, OWNER
- Description: These fields provide location details. BORO Indicates the borough, with fully populated and with 5 unique values (e.g., Manhattan = 1, Bronx = 2). STADDR represents street addresses, with 839,280 unique values and 99.9% population. ZIP contains ZIP codes, with 196 unique values and 97.2% population. OWNER indicates the owner of the property, 97% populated with 863,347 unique values. The most common owner is "PARKCHESTER PRESERVAT," indicating potential shared ownership.

2.3) Building and Property Attributes

- Fields: EASEMENT, BLDGCL, TAXCLASS, EXT
- Description: These provide classification and details about the building or property. EASEMENT indicates easement types, with 12 unique values. Only 0.4% of records are populated. BLDGCL represents the building class, with 200 unique values and fully populated. TAXCLASS indicates tax classification, with 11 unique values and 100% population. Most common value is 1 (residential properties). EXT indicates extension types, with 33.1% population and 3 unique values.

2.4) Exemption and Adjustment Codes

- Fields: EXCD1, EXCD2, EXMPTCL
- Description: They capture exemption or adjustment codes for properties. EXCD1 and EXCD2 are exemption codes, with 59.6% and 8.7% population, respectively. Both fields share 1017 as the most common value. EXMPTCL indicates exemption classes, with 14 unique values. Only 1.5% of records are populated, with X1 as the most common value.

2.5) Assessment and Value Type Information

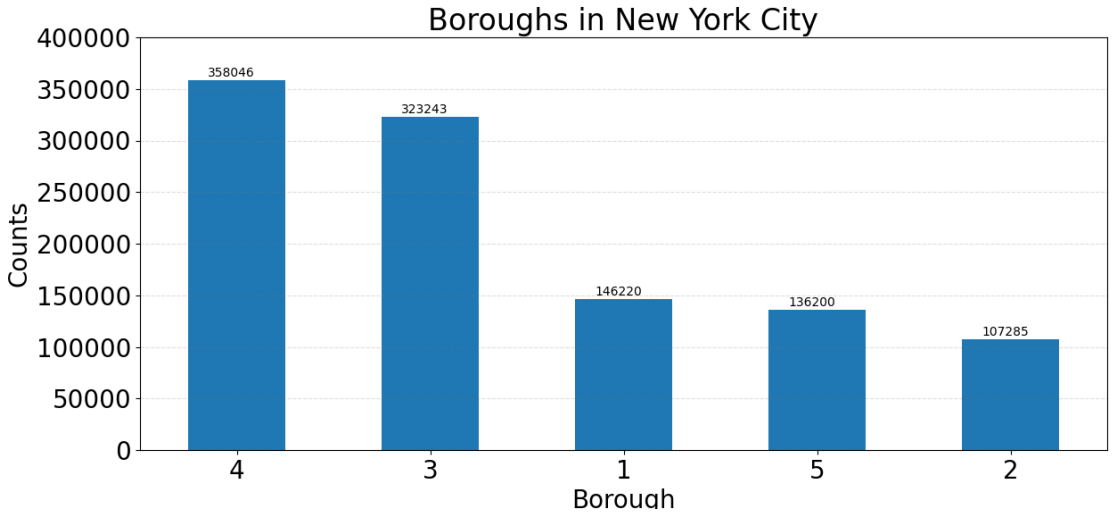
- Fields: PERIOD, YEAR, VALTYPE
- Description: These fields provide metadata related to the assessment and valuation process. PERIOD indicates the assessment period when the data was created and it is fully populated with 1 unique value (FINAL). YEAR represents the assessment year and it is fully populated with 1 unique value (2010/11). VALTYPE specifies the valuation type used during the assessment and it is fully populated with 1 unique value (AC-TR).

3. Distribution

Some fields were exempted as they are either completely unique or identical across all records. RECORD and BBLE are unique identifiers, offering no meaningful patterns for analysis. PERIOD, YEAR, and VALTYPE contain a single identical value, providing no insights for visualization.

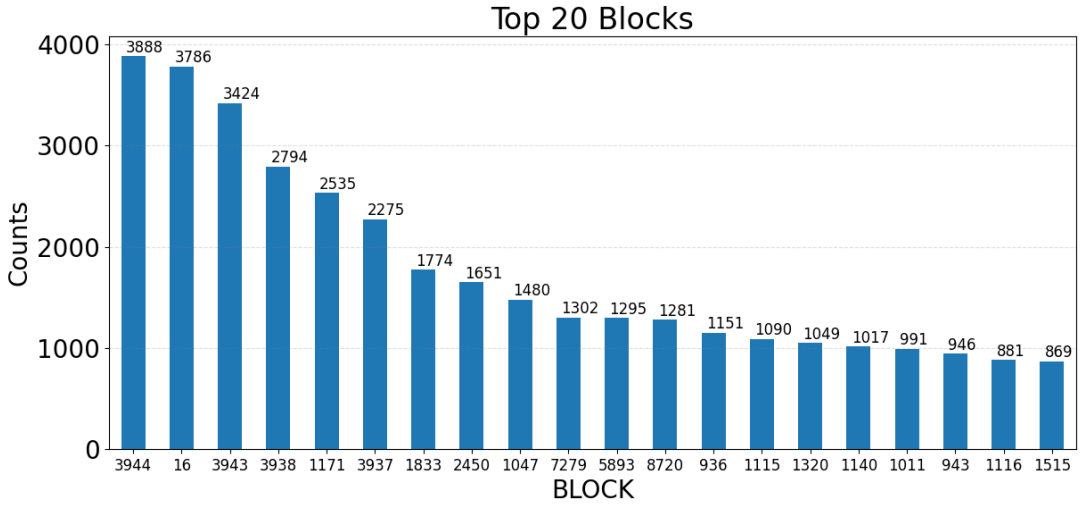
1) Boroughs (BORO)

The chart shows the distribution of property records across New York City's boroughs. Queens (358,046) and Brooklyn (323,243) have the most records, followed by Manhattan (146,220) and Staten Island (136,200). The Bronx has the fewest records (107,285). This reflects an uneven distribution of property records across the boroughs.



2) Block (BLOCK)

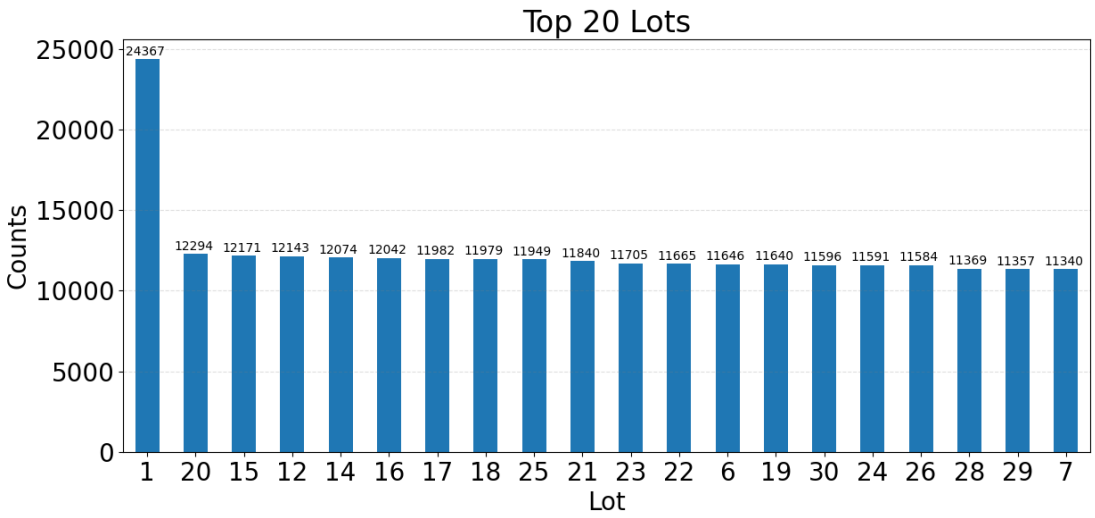
The chart highlights blocks with disproportionately high numbers of property records, such as 3944 (3,888) and 16 (3,786). In the context of identifying potential property tax fraud, these blocks may warrant closer inspection for patterns that deviate from norms.

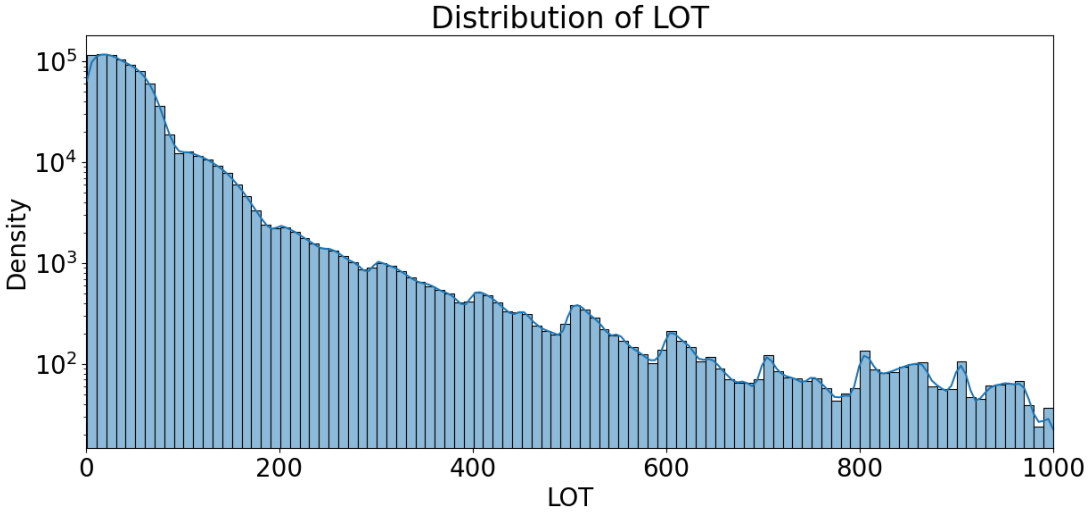


3) LOT (LOT)

Lot 1 has a significantly higher count (24,367), far exceeding the other top lots. This suggests that Lot 1 is a common identifier, potentially representing a standard or default lot designation. The remaining lots have counts that are relatively close to each other, ranging between approximately 11,340 and 12,294.

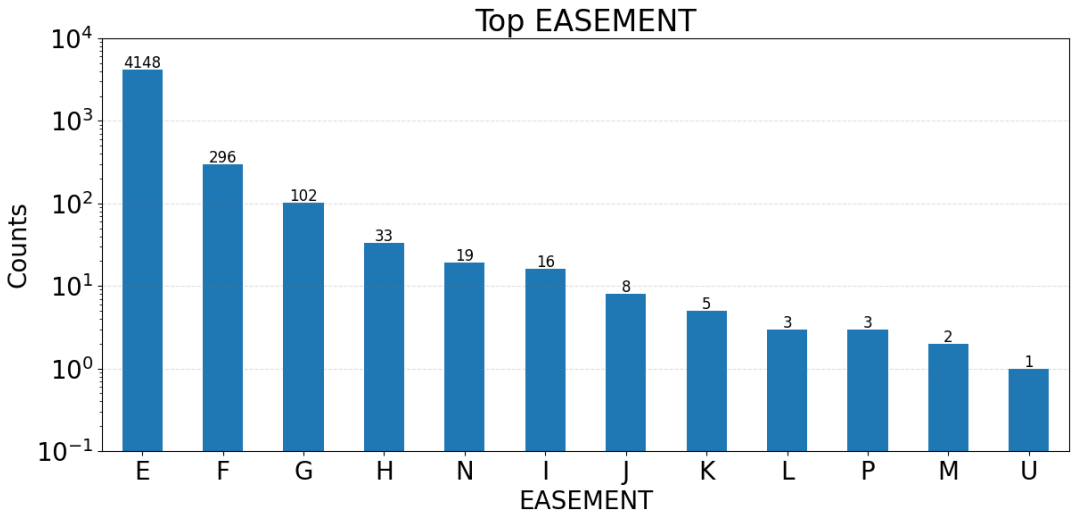
The second chart shows the distribution of LOT values limited to <=1000. The density is significantly higher at lower LOT values (e.g., <200) and declines sharply as values increase. This pattern suggests that most properties are concentrated in smaller lot numbers, likely representing standard residential or small-scale properties. These lower LOT values may correspond to high-density residential areas, where tax exemptions or misreporting could be more prevalent.





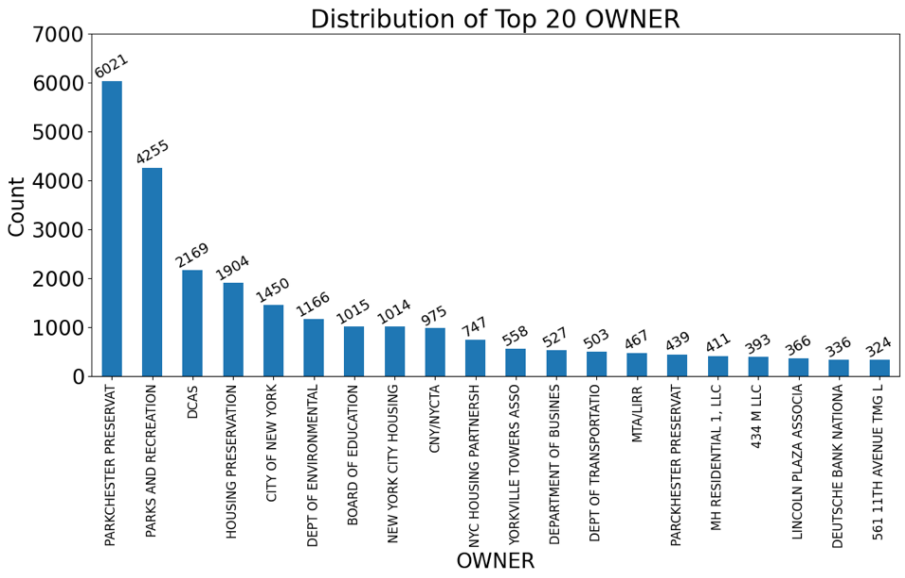
4) Easement (EASEMENT)

The easement chart shows that easement type E dominates with 4,148 occurrences, far exceeding other types, such as F (296) and G (102). This skewed distribution suggests that most properties fall under standard easement categories, with limited variability.



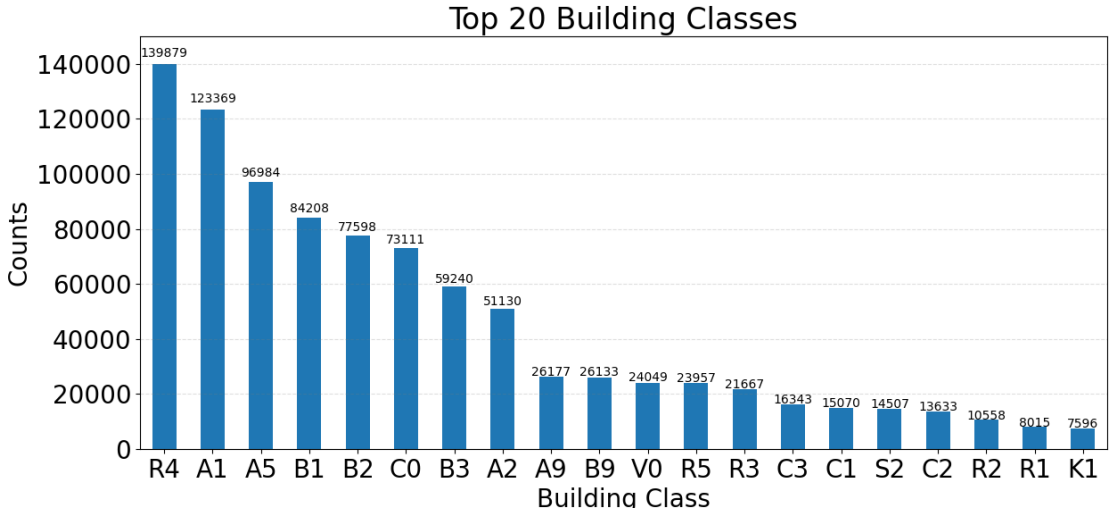
5) Owner (OWNER)

The owner chart highlights the concentration of ownership among a few entities, with the top owner, "PARKCHESTER PRESERVAT," holding 6,021 properties, followed by "RECREATION" (4,255) and "DCAS" (2,169). This indicates that a significant portion of properties is owned by organizations or large-scale property holders.



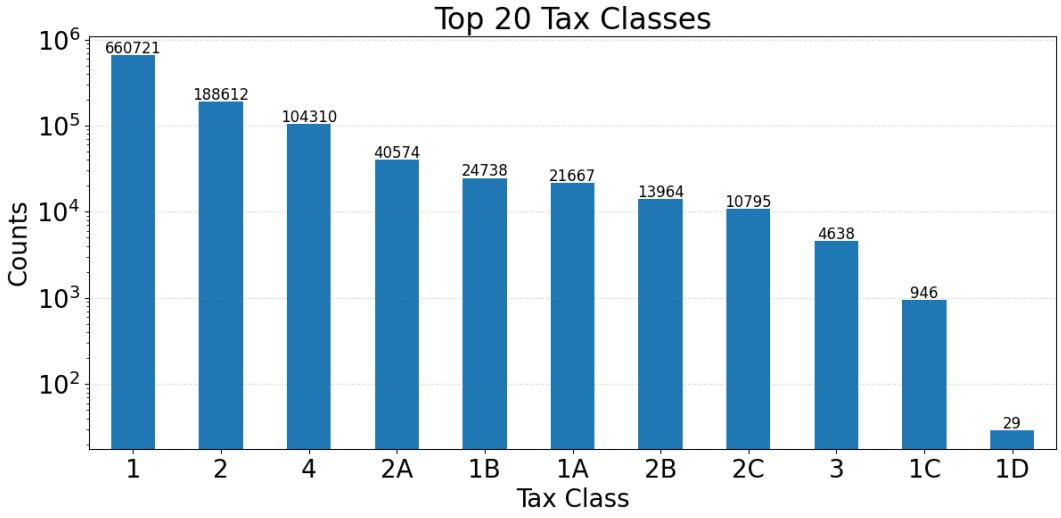
6) Building Class (BLDGCL)

The chart displays the top 20 building classes, with R4 dominating at 139,879 properties, followed by A1 (123,369) and A5 (96,984). These top classes likely represent common residential property types, such as single-family or multifamily homes. Lower counts for other classes, such as K1 (7,596), may represent specialized or less frequent building types like industrial, mixed-use, or commercial properties.



7) Tax Class (TAXCLASS)

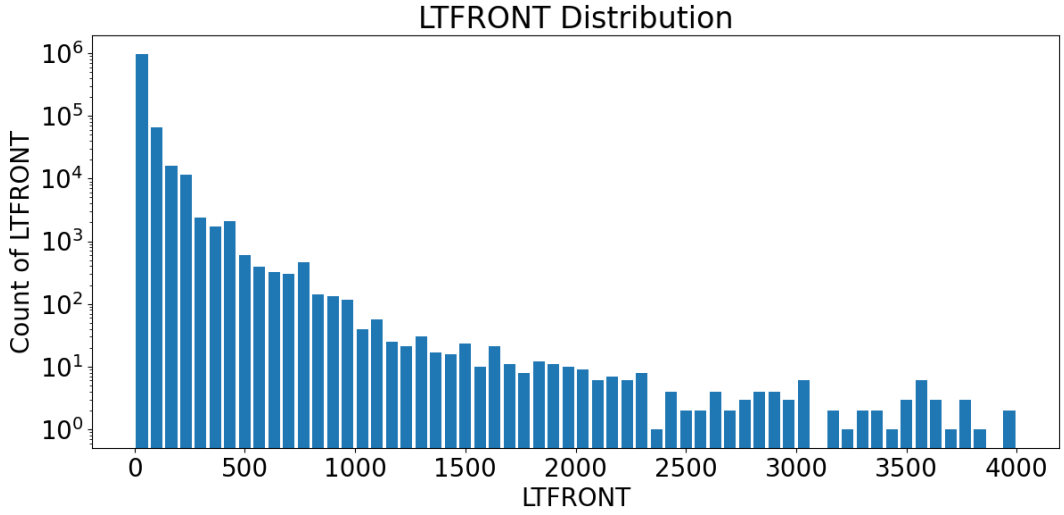
The chart for tax classes shows that 1 is overwhelmingly the most common, with 660,721 records, followed by 2 and 4. Tax class 1 is typically associated with residential properties (e.g., single-family homes), while 2 and 4 often represent multifamily residential and commercial properties, respectively. The sharp dominance of class 1, 2, and 4 suggests that the dataset is heavily skewed toward residential properties. Rarer classes, such as 1D (29 records), could represent specialized property types or administrative designations, requiring closer scrutiny.

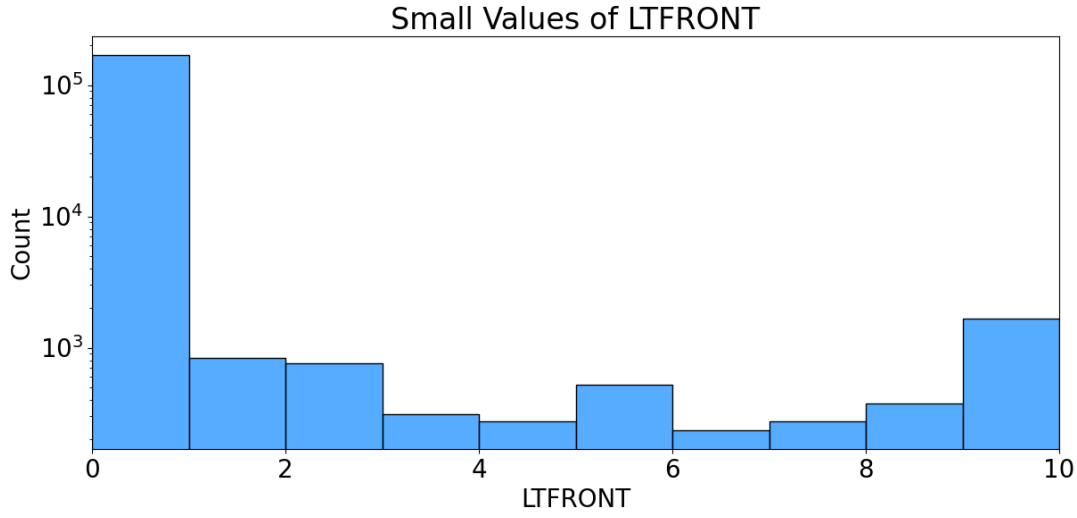


8) Lot Width (LTFRONT)

The first chart displays the overall distribution of lot width (LTFRONT), highlighting a highly skewed pattern with the majority of properties below 500 feet. This indicates that most lots in the dataset are relatively narrow, typical of urban residential or small-scale properties, while larger frontages are less frequent and likely represent commercial or industrial properties.

The second chart focuses on frontage values between 0 and 10 feet, showing a significant spike at 0 feet. This could indicate missing or unmeasured data for certain properties. Smaller non-zero values, such as those under 5 feet, are rare and may represent unusually narrow or irregularly shaped lots, often found in densely populated urban areas.

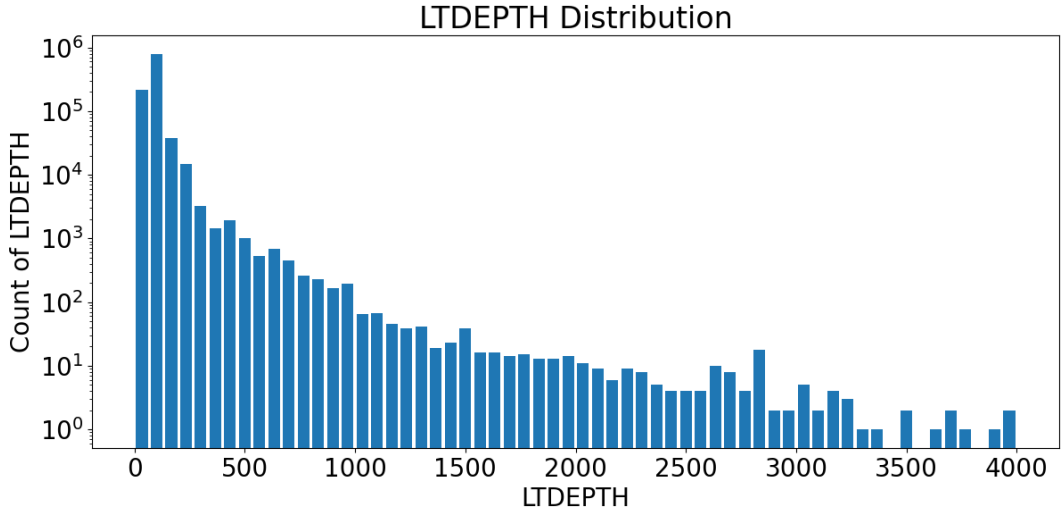


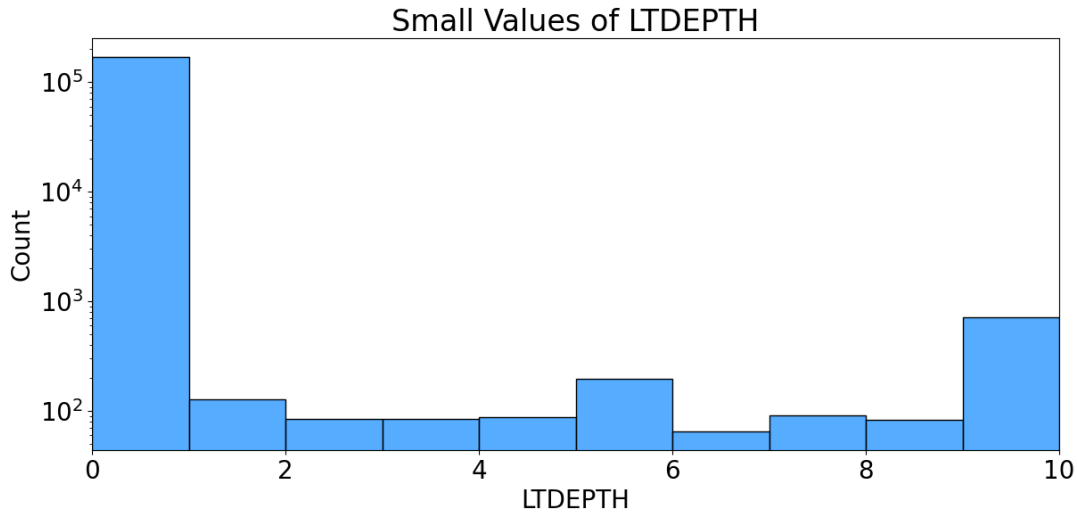


9) Lot Depth (LTDEPTH)

The first chart shows the overall distribution of LTDEPTH, with a highly skewed pattern. The majority of properties have lot depths below 500 feet, suggesting that most lots are relatively shallow, typical of urban residential or small-scale commercial properties. As lot depth increases, the frequency decreases significantly, with very few properties exceeding 1000 feet. These larger lot depths are likely associated with specialized uses such as industrial, agricultural, or large commercial properties.

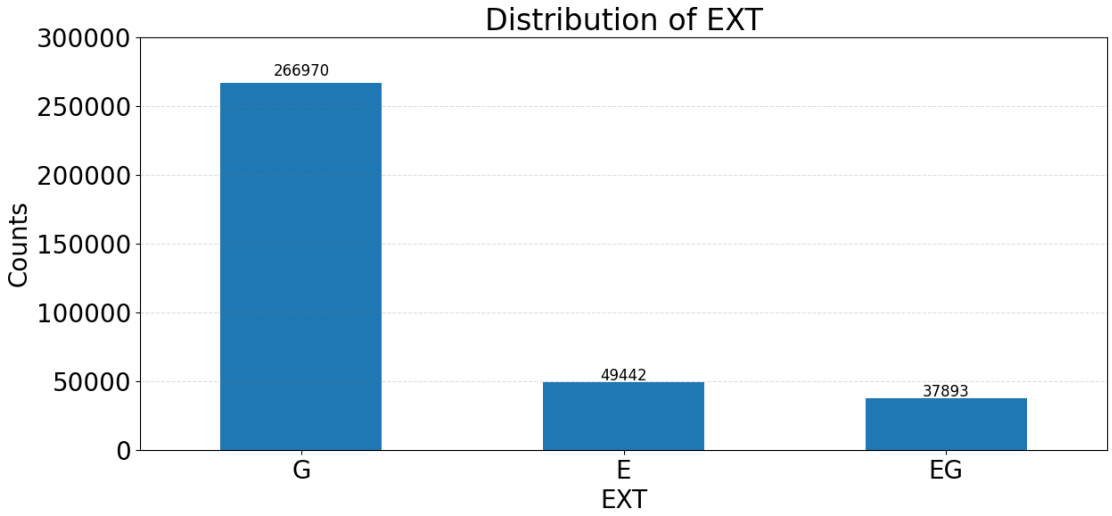
The second chart zooms in on small LTDEPTH values between 0 and 10 feet. A substantial spike is observed at 0 feet, likely representing missing or unmeasured data. Other small non-zero values (e.g., 1-10 feet) are rare but could represent irregular or exceptionally small lots, possibly due to unique zoning configurations or errors in data entry.





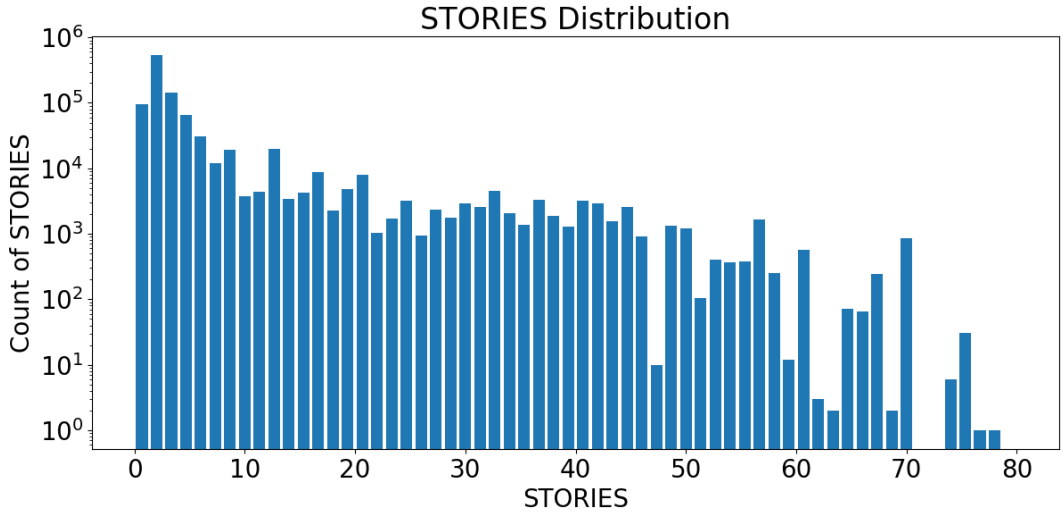
10) Extension (EXT)

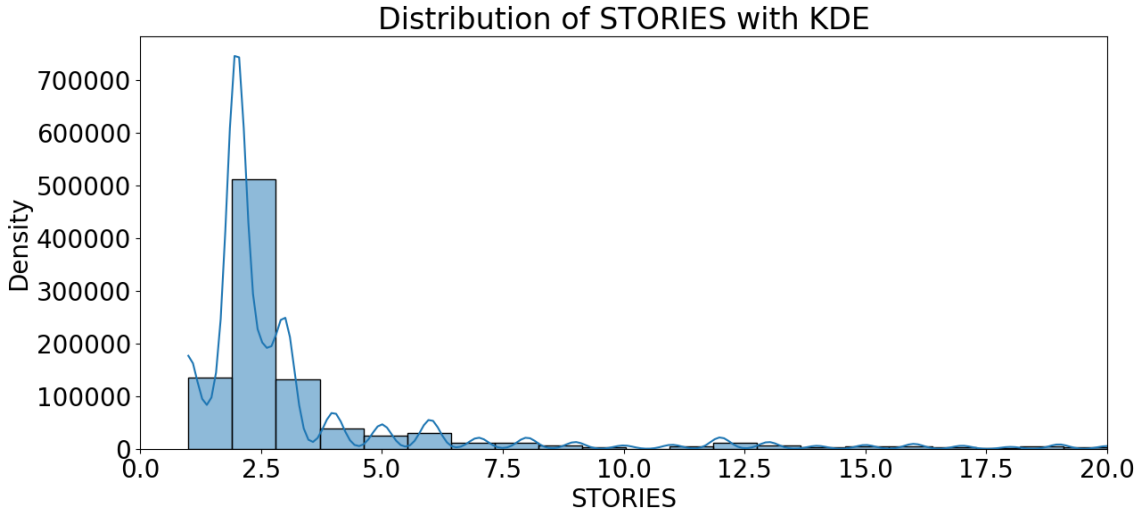
The EXT field indicates the type of extension applied to a property. The most common value is 'G' with over 260,000 occurrences, suggesting that this extension type is widely implemented or represents a default classification. Other extension types, such as 'E' (49,442) and 'EG' (37,893), occur far less frequently, indicating that they are applied to a smaller subset of properties.



11) Stories (STORIES)

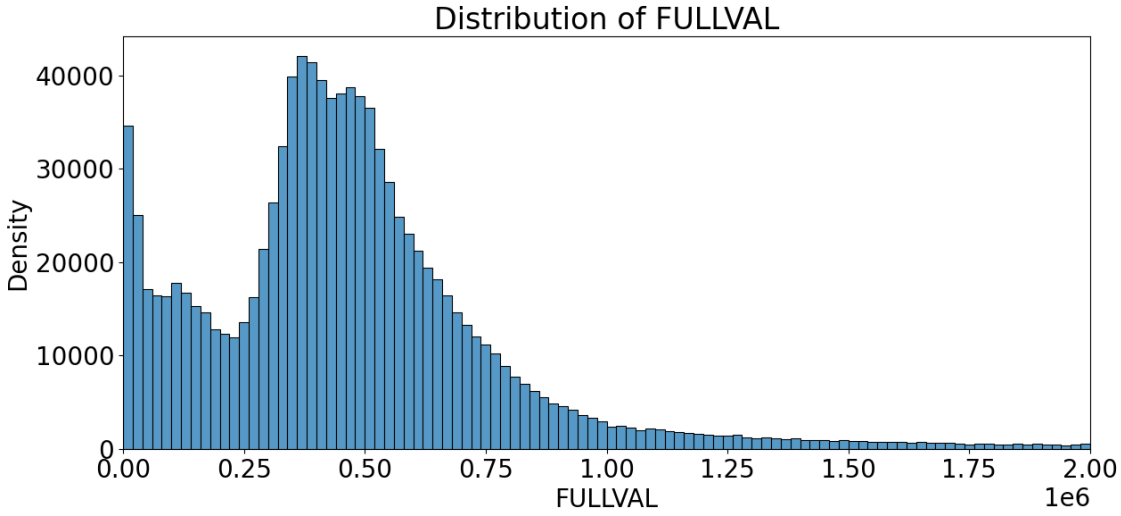
The STORIES field represents the number of floors in a building. The distribution is skewed toward lower values, with the majority of properties having fewer than 10 stories, indicative of typical residential or small commercial buildings. The frequency decreases steadily as the number of stories increases, with a few exceptions around higher values, likely representing high-rise buildings or specialized structures. A closer examination with a KDE plot highlights that the most common number of stories is around 2-3, with a rapid decline in density as the number of stories increases.

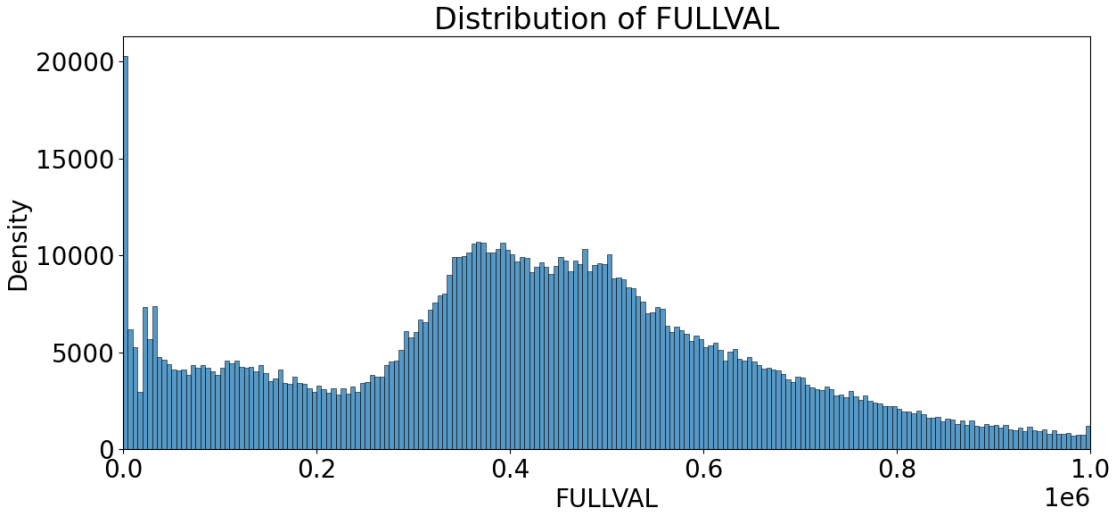




12) Market Value (FULLVAL)

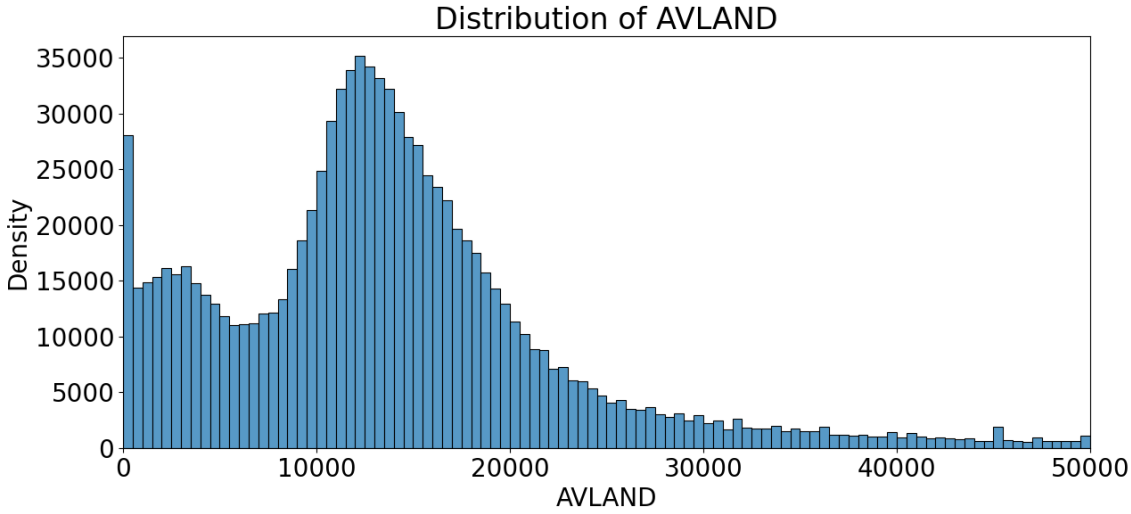
The distribution is right-skewed, indicating that the majority of properties are valued below 1 million, with a prominent density peak around the $250,000–$500,000 range. This reflects the prevalence of mid-range residential properties in the dataset. The second graph focuses on a narrower range and highlights a clustering of properties between $200,000 and $600,000, which is consistent with typical urban residential markets.





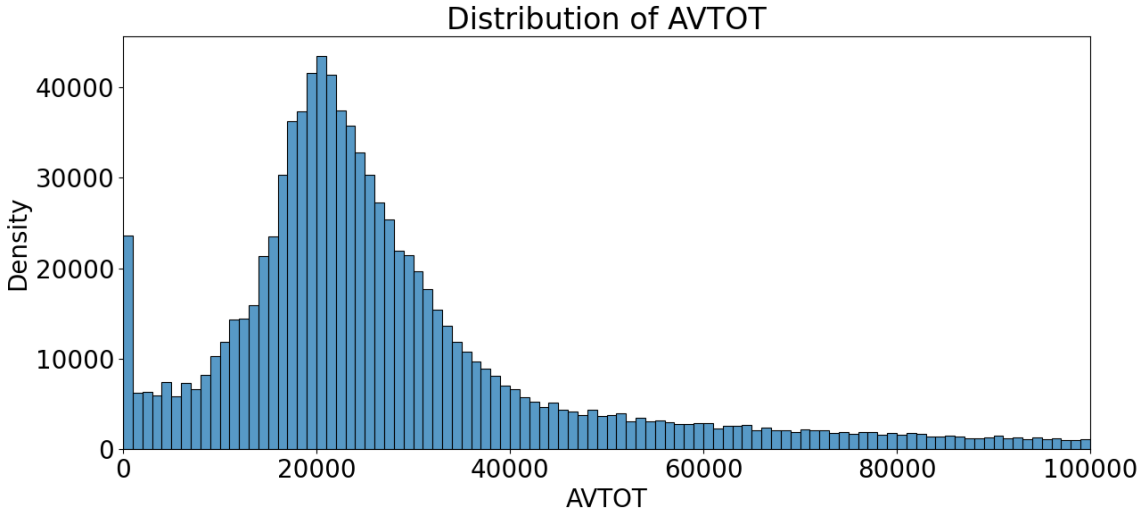
13) Actual Land Value (AVLAND)

The distribution of AVLAND (Actual Land Value) is right-skewed, with the majority of properties concentrated in the $10,000–$20,000 range. There is a distinct peak around $15,000, suggesting that a large proportion of properties are assigned mid-range land values, typical for residential or small commercial lots in urban areas. The distribution gradually tapers off beyond $20,000, reflecting fewer high-value properties, likely associated with larger or premium land parcels.



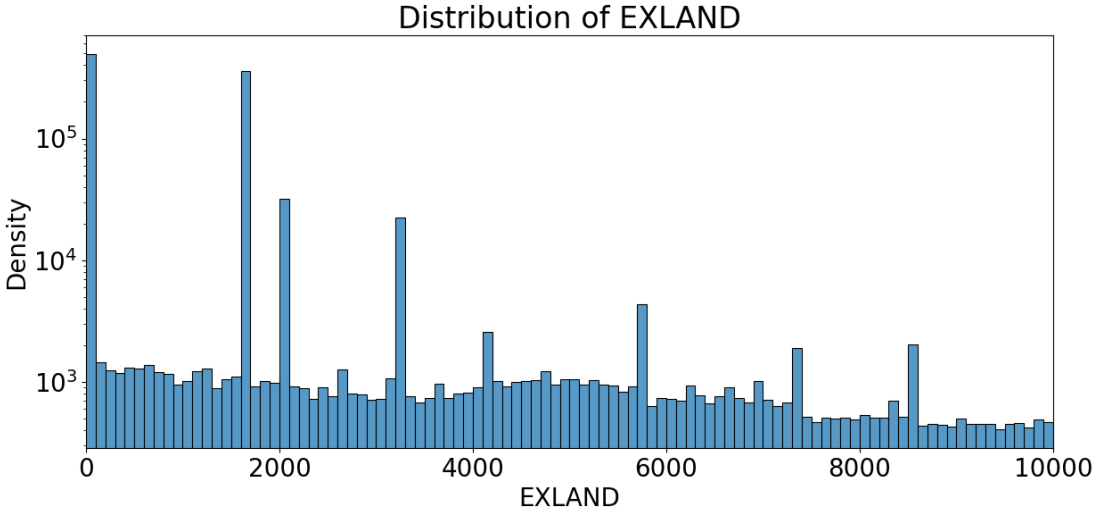
14) Actual Total Value (AVTOT)

The distribution of AVTOT is right-skewed, with the majority of properties concentrated between $20,000 and $30,000. This suggests many properties are assessed within this range, representing typical combined land and building values for residential or small commercial properties. The tail extends to higher values, reflecting fewer properties with significantly higher total valuations, such as large commercial, industrial, or premium properties.



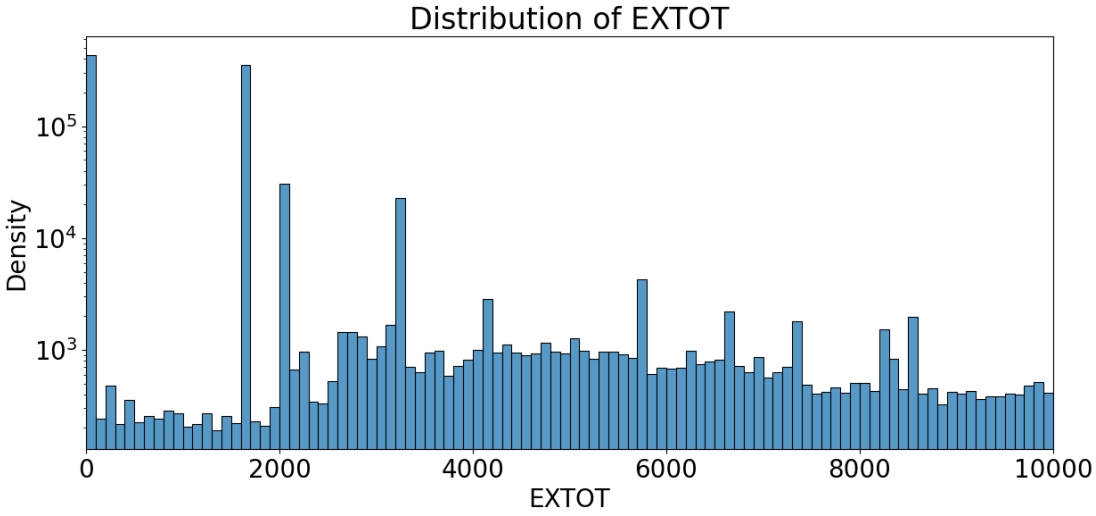
15) Actual Exempt Land Value (EXLAND)

The distribution of EXLAND is sparse and exhibits notable peaks at specific values, such as around $1,000, $2,000, and $5,000. This pattern suggests that exempt land values are concentrated at certain standardized amounts, possibly due to common tax exemption brackets or policies. The high density near zero indicates that many properties have little to no exempt land value, reflecting ineligibility for exemptions.



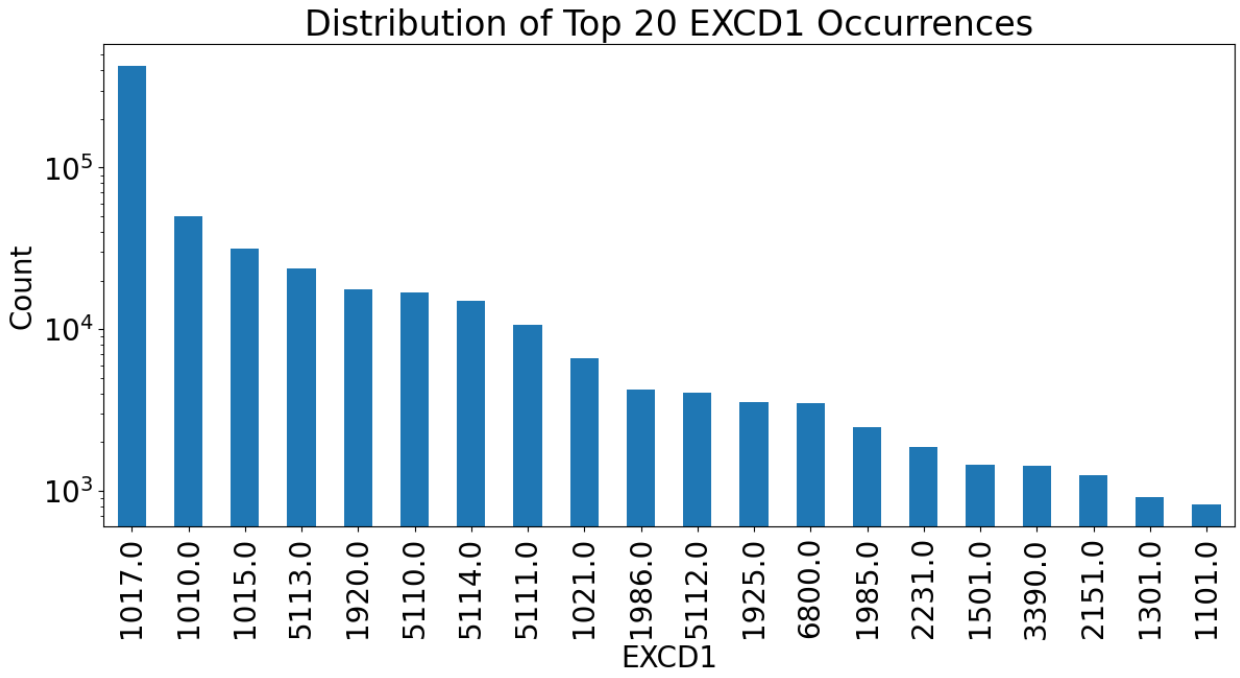
16) Actual Exempt Total Value (EXTOT)

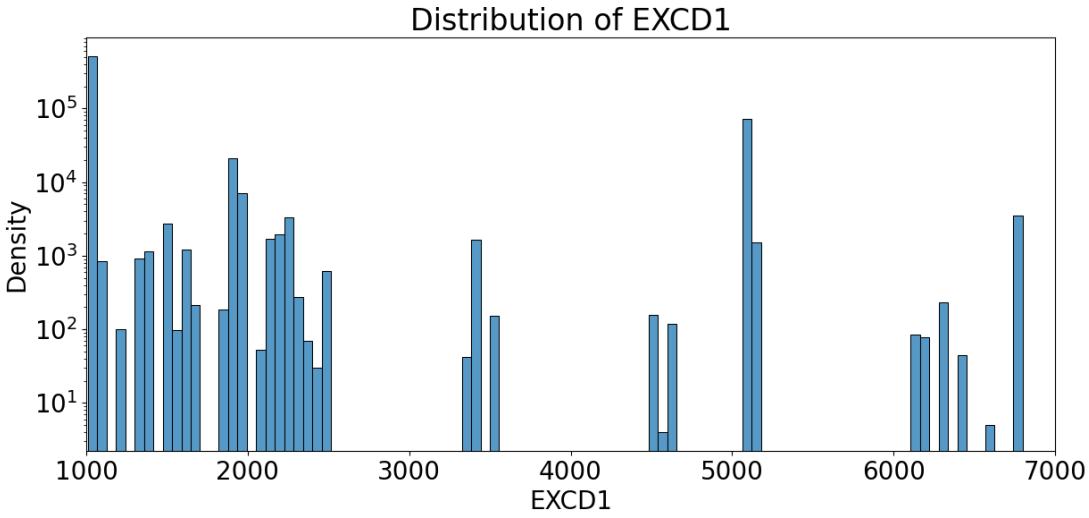
The distribution of EXTOT exhibits similar characteristics to EXLAND, with pronounced peaks at specific values such as $1,000, $2,000, and $5,000, indicating standard exemption amounts applied to a significant number of properties. The majority of records cluster around low exemption values, with a high density near zero, suggesting that many properties have minimal or no total exemptions.



17) Exemption Code 1 (EXCD1)

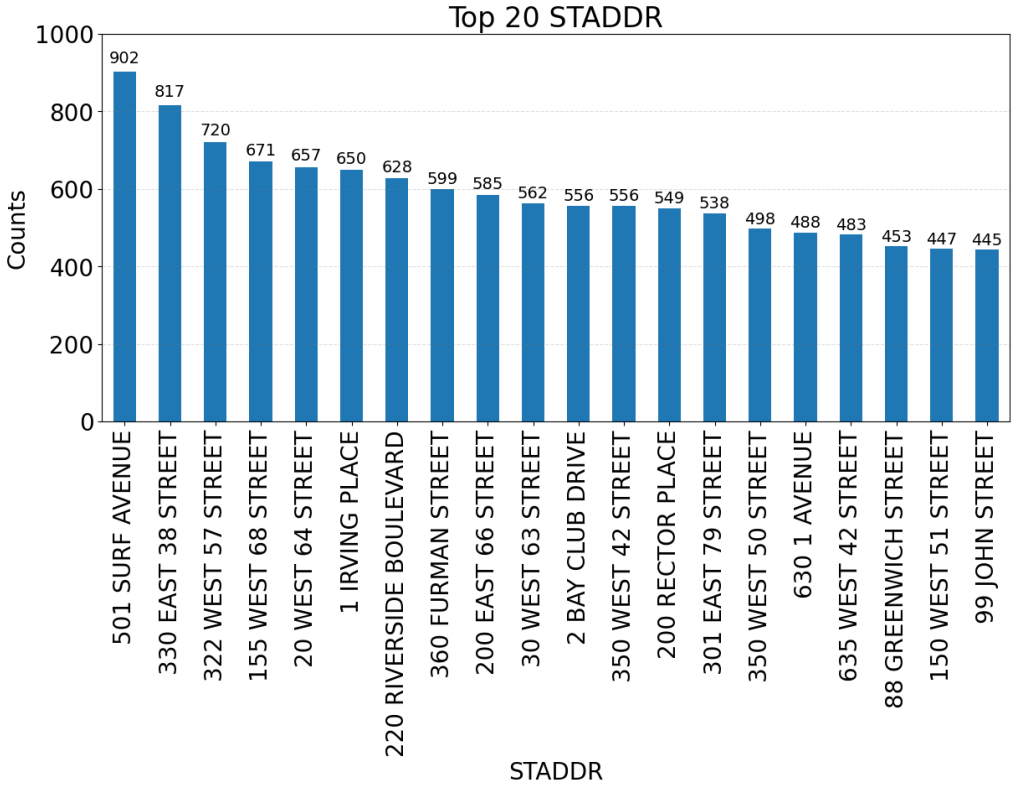
The EXCD1 field distribution reveals a clear hierarchy in the frequency of exemption codes. The most common code, 1017.0, appears significantly more frequently than others, suggesting it represents a widely applicable exemption category. Other frequent codes, such as 1010.0 and 1015.0, are also prominent, though they occur less often than the leading code. The second plot indicates that exemptions are not distributed evenly but are concentrated in certain categories, likely tied to specific property types or owner characteristics. The systematic distribution highlights the structured application of tax exemptions, with specific codes representing standardized policies.





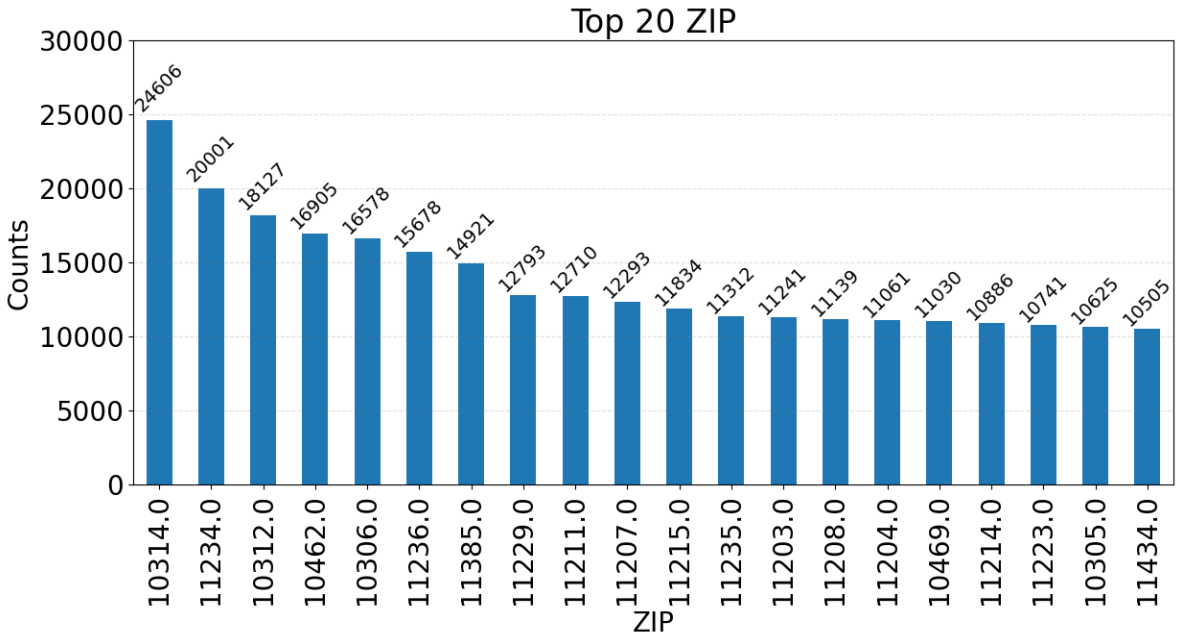
18) Street Address (STADDR)

The STADDR field, which represents the street address of properties, shows a highly uneven distribution. The most frequently occurring address, 501 Surf Avenue, appears 902 times, significantly more than others. Other addresses, such as 330 East 39 Street and 322 West 57 Street, also exhibit high counts, though they trail behind the top address.



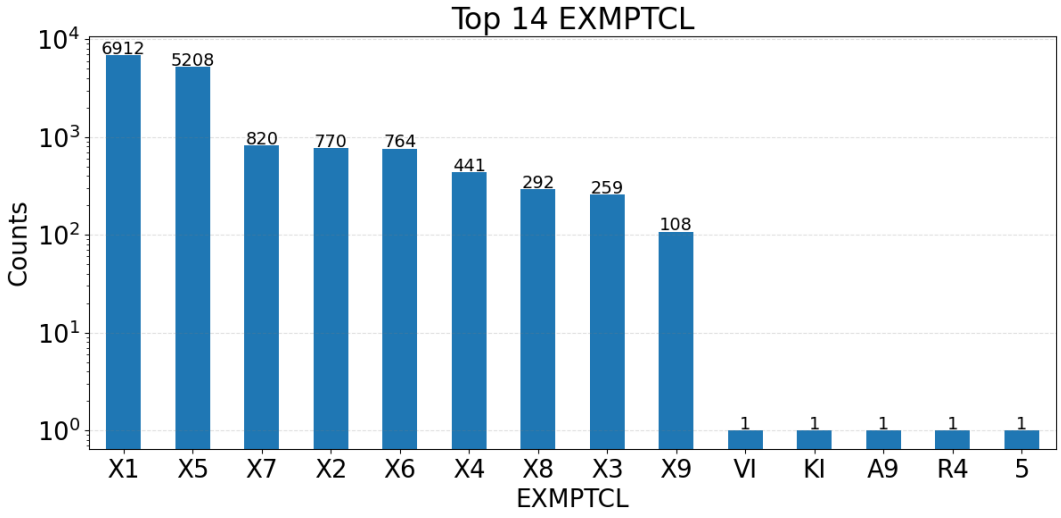
19) Zip Code (ZIP)

The ZIP field, representing property zip codes, shows notable variation in counts across different areas. This pattern suggests that certain areas, such as Staten Island (10314.0) and Brooklyn (11234.0, 11236.0), have higher concentrations of properties in the dataset. These areas might correspond to large residential zones, multi-unit properties, or specific urban planning characteristics.



20) Exemption Class (EXMPTCL)

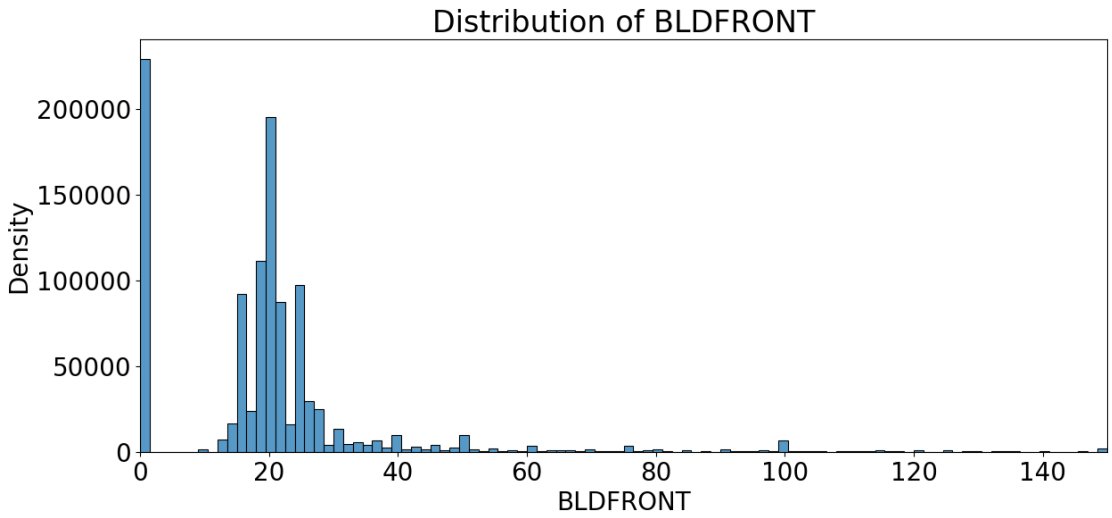
The EXMPTCL field, representing exemption classes, comprises 14 unique values. The concentration of exemptions in specific classes (e.g., X1 and X5) likely reflects prevalent tax policies or property characteristics associated with those categories. The rare exemption classes may correspond to specialized exemptions or unique property types.

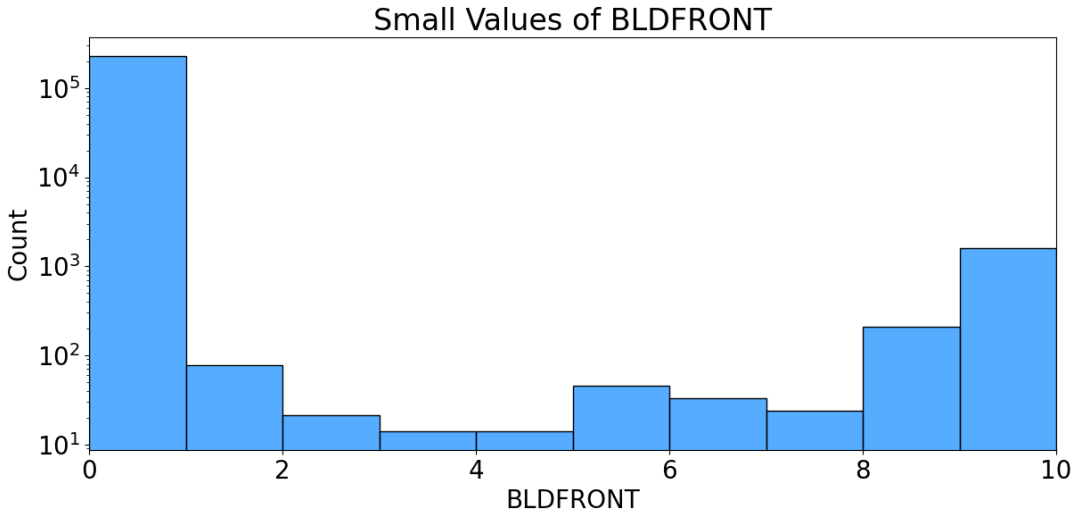


21) Building Width (BLDFRONT)

The BLDFRONT field, representing building widths, exhibits a heavily skewed distribution with most properties concentrated below 40 feet. A clear peak is observed around 20–25 feet, suggesting that this is the standard width for many urban residential or small commercial buildings. Outliers exist with widths exceeding 100 feet, corresponding to large commercial or industrial properties.

The second chart highlights small building widths (≤10 feet), with a significant spike at 0 feet. Non-zero small widths, such as 8–10 feet, may correspond to narrow or irregularly shaped structures commonly found in densely populated urban areas. The prevalence of 0-width records and unusually narrow properties warrants further examination to identify possible anomalies.

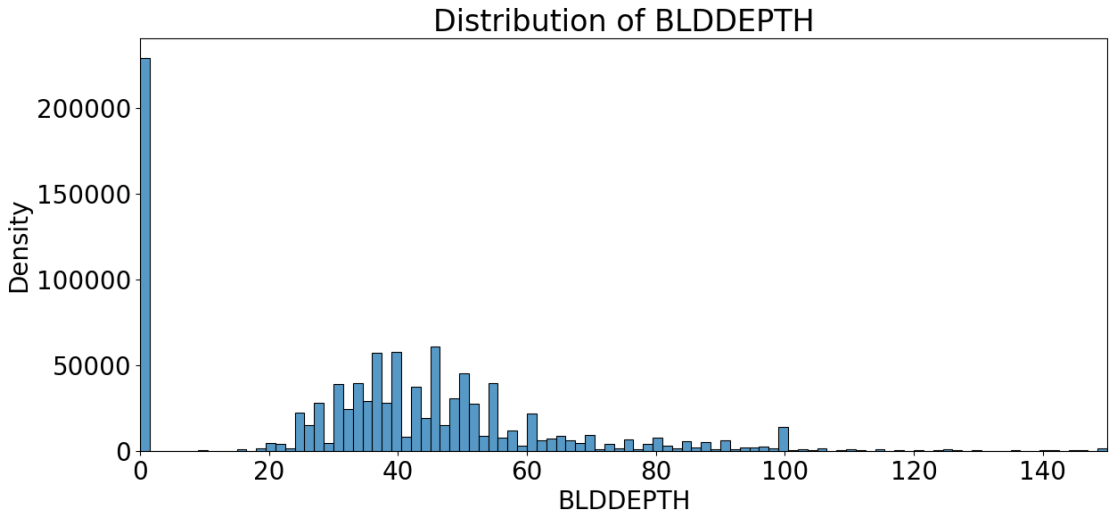


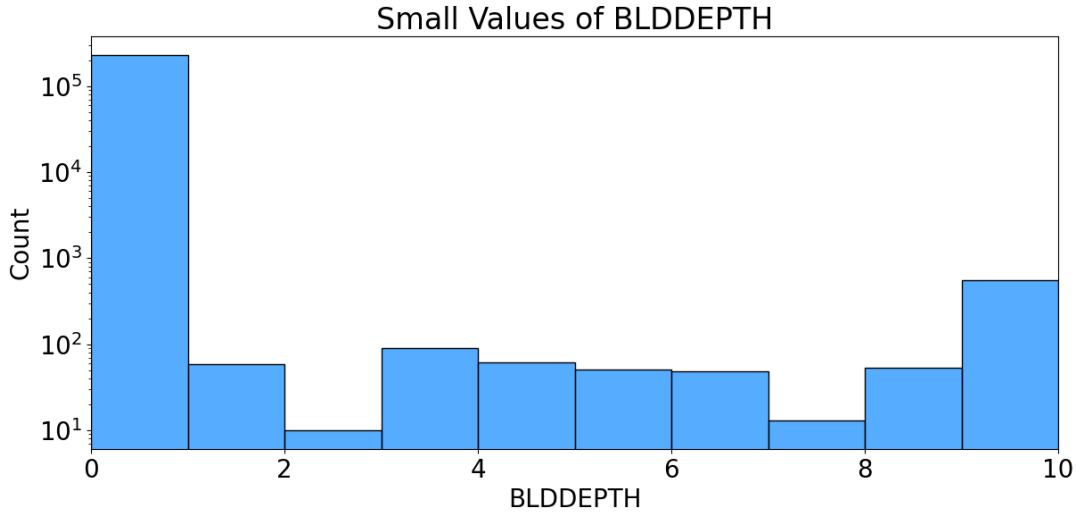


22) Building Depth (BLDDEPTH)

The BLDDEPTH field, representing building depths, shows a skewed distribution with most properties concentrated below 60 feet. A significant spike at 0 suggest a notable portion of properties have missing or unmeasured depth values. Also, a distinct peak is visible around 40–50 feet, indicating a standard depth for many residential or small commercial buildings.

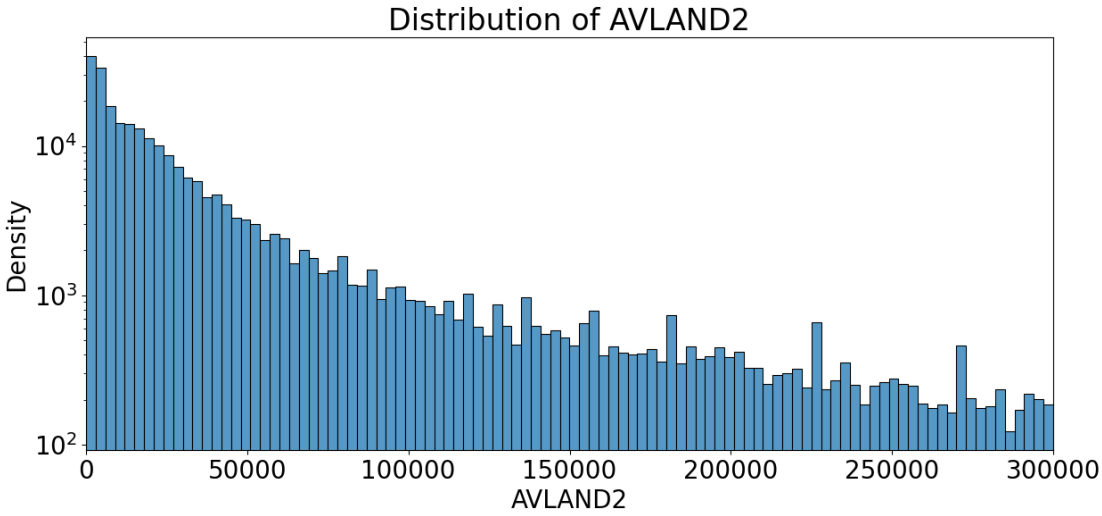
The second chart focuses on small building depths (≤10 feet), with a significant spike at 0 feet, suggesting missing or unrecorded data. Non-zero small depths (e.g., 8–10 feet) may indicate narrow or irregularly shaped buildings, potentially in densely populated urban areas.





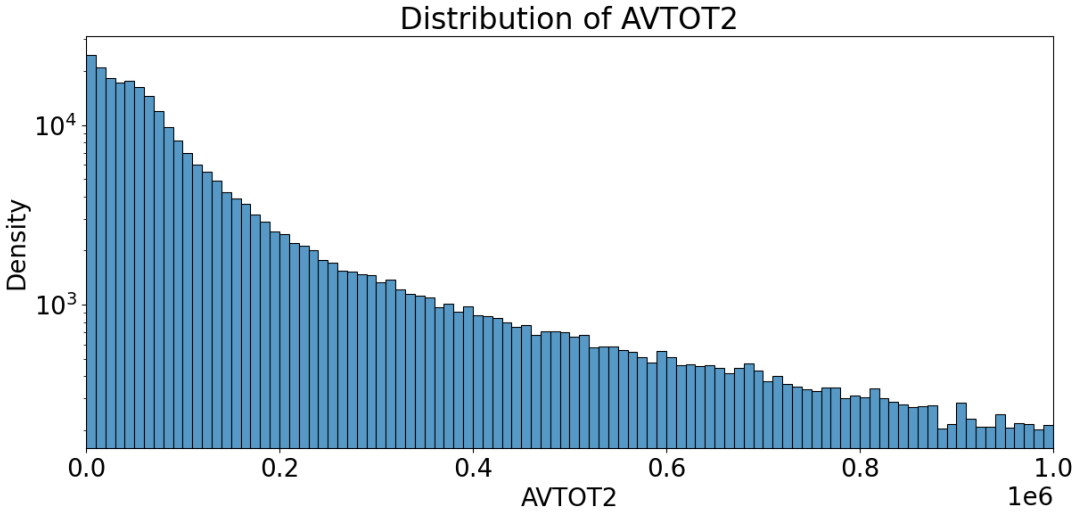
23) Transitional Land Value (AVLAND2)

The distribution of AVLAND2 is right-skewed, with the majority of properties concentrated at lower transitional land values, declining steadily as values increase. The high density near the lower range suggests that most properties have relatively modest transitional land valuations, likely reflecting smaller or less complex properties.



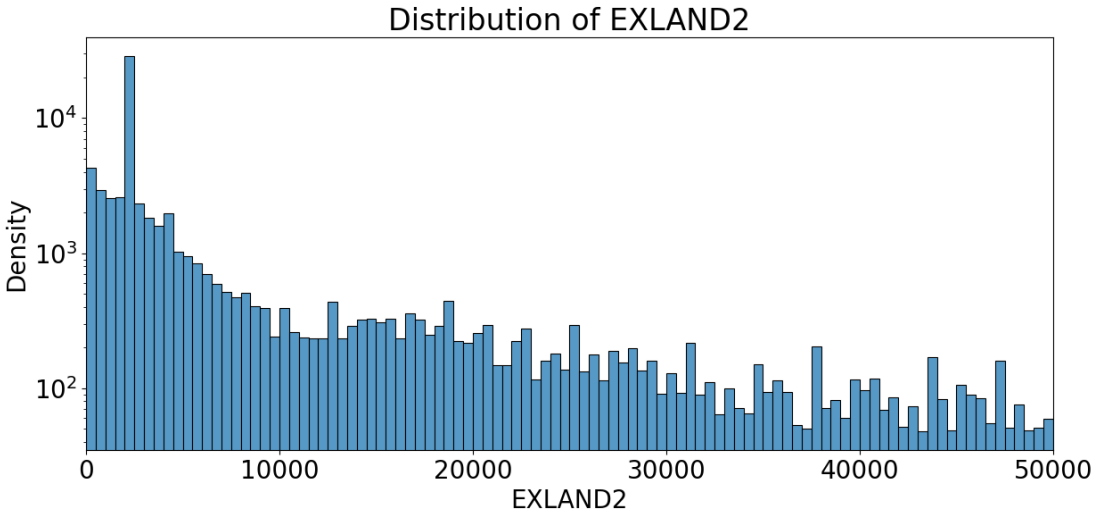
24) Transitional Total Value (AVTOT2)

The distribution of AVTOT2 similarly exhibits a strong right skew, with most properties clustered at lower total values and a gradual decline in density as values increase. This pattern reflects the dominance of properties with relatively low transitional total valuations, while the small number of higher-value properties could include large commercial buildings or high-value residential properties that may be influenced by transitional assessments or exemptions.



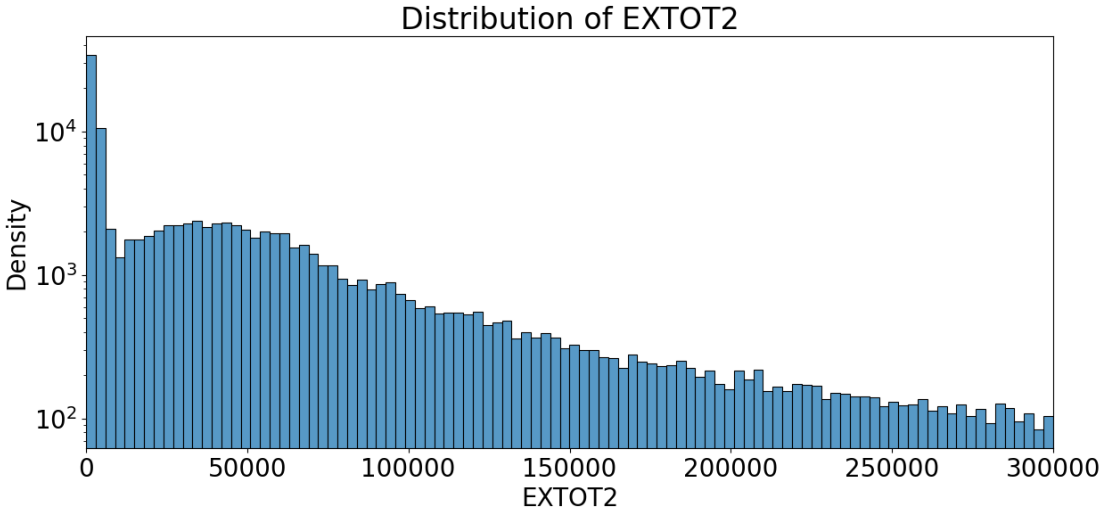
25) Transitional Exemption Land Value (EXLAND2)

The distribution of EXLAND2 indicates a steep decline from a high concentration of low values, with a long tail representing properties with higher transitional exemption land values. The spike at 0 suggests that many properties have no recorded exemption land value, while the range of values up to 50,000 indicates a minority of properties with significant exemptions.



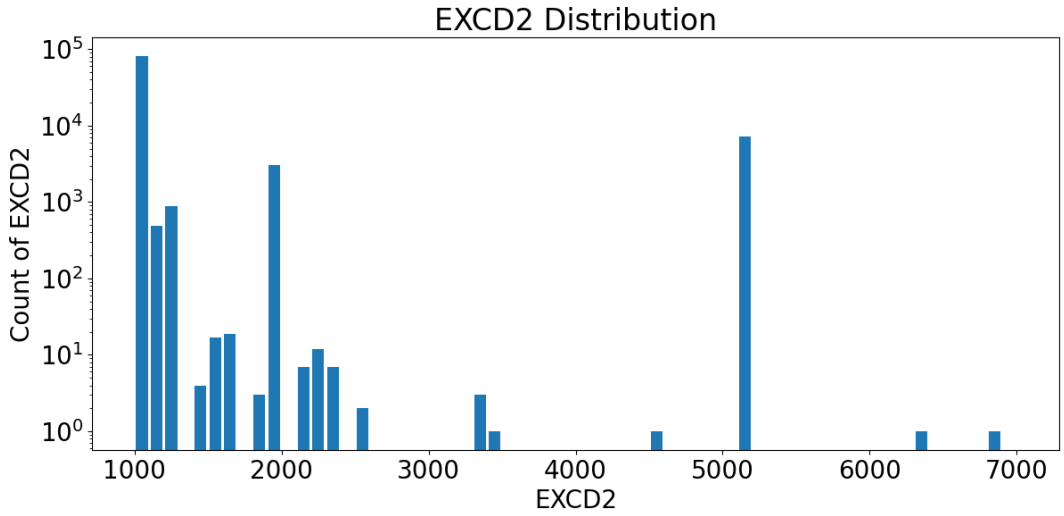
26) Transitional Exemption Land Total Value (EXTOT2)

The EXTOT2 distribution reflects a similar pattern to EXLAND2, with a steep initial drop-off and a long right tail. The values range up to 300,000, with a significant cluster of properties near 0. The higher total exemption values may correspond to properties with both land and building exemptions, influenced by transitional tax policies or large-scale exemptions.



27) Exemption Code 2 (EXCD2)

The distribution of EXCD2 shows distinct clusters of values, with certain codes like 1000, 2000, and 5000 occurring more frequently. The concentration in these categories indicates specific types of exemptions applied across properties. The less frequent but distinct higher values may represent specialized exemptions, potentially tied to unique use cases or tax statuses.



III. Data Cleaning

1. Purpose

The purpose of this data cleaning process is to ensure a high-quality dataset for unsupervised fraud detection by retaining outliers as potential anomalies, excluding irrelevant public or government-owned records to focus on private property transactions, and systematically imputing missing values to maintain data completeness and reliability.

2. Process

1) Outliers

In unsupervised fraud detection, outliers are not removed since they represent potential instances of fraudulent activity, which is the focus of the analysis. Instead, the outliers are preserved and investigated as they indicate unusual patterns or anomalies that align with fraudulent behavior.

2) Exclusions

To focus on identifying tax fraud by private property owners, publicly or government-owned properties were removed. The process filtered out records with easement type "U" and identifies potential government or public owners using keywords (e.g., "GOVERNMENT," "DEPARTMENT") and frequently occurring owner names. Specific public or governmental owner names were manually added to the exclusion list for thorough filtering. A total of 26,502 records were removed, ensuring that the dataset prioritizes private properties relevant to the investigation.

3) Imputations for Missing Values

- **Zip code (ZIP)**
  - There were 20,431 missing values in the Zip code field.
  - A new column, staddr_boro, was created by concatenating STADDR (street address) and BORO (borough) for records where both fields were non-null. Using this combined field, a dictionary was constructed to map staddr_boro to known ZIP codes, and missing values were filled through this mapping process (2,832 filled, 17,599 remaining).
  - Forward and backward fill methods were applied, checking if the preceding and succeeding records had the same ZIP code. If they matched, the missing ZIP was filled with that value (16,126 filled, 1,473 remaining).
  - The remaining missing ZIP codes were resolved using the forward fill method alone, filling them with the ZIP code from the previous record (1,473 filled, 0 remaining).

- **Market Value (FULLVAL)**
  - There were 10,025 missing values in the FULLVAL.
  - The missing values were replaced with NaN for easier handling.
  - The field was grouped by TAXCLASS, BORO, and BLDGCL, and the missing values were filled with the group mean (2,718 filled, 7,307 remaining).
  - Next, the field was grouped by TAXCLASS and BORO, and the remaining missing values were filled with the group mean at this level (6,921 filled, 386 remaining).

- Finally, the field was grouped by TAXCLASS alone, and the remaining missing values were filled with the group mean at this level (386 filled, 0 remaining).

- **Actual Land Value (AVLAND)**
  - There were 10,027 missing values in the AVLAND.
  - The missing values were replaced with NaN for easier handling.
  - The field was grouped by TAXCLASS, BORO, and BLDGCL, and the missing values were filled with the group mean (2,720 filled, 7,307 remaining).
  - Then, grouping by TAXCLASS and BORO, the missing values were filled with the group mean at this level (6,921 filled, 386 remaining).
  - Finally, the remaining missing values were filled with the group mean based on TAXCLASS alone (386 filled, 0 remaining).

- **Actual Total Value (AVTOT)**
  - There were 10,025 missing values in the AVTOT.
  - The missing values were replaced with NaN for easier handling.
  - The field was grouped by TAXCLASS, BORO, and BLDGCL, the missing values were filled with the group mean (2,718 filled, 7,307 remaining).
  - The field was then grouped by TAXCLASS and BORO, and the remaining missing values were filled with the group mean at this level (6,921 filled, 386 remaining).
  - Finally, the remaining missing values were filled with the group mean at the TAXCLASS level (386 filled, 0 remaining).

- **Stories (STORIES)**
  - There were 42,029 missing values in the STORIES.
  - The data was first grouped by BORO and BLDGCL, and missing values were filled using the most frequently occurring value within each group (4,108 filled, 37,921 remaining).
  - Next, the data was grouped by TAXCLASS, and the remaining missing values were filled using the mean of STORIES within each group (37,921 filled, 0 remaining).

- **Lot Width (LTFRONT)**
  - There were 161,133 missing values in the LTFRONT.
  - The invalid values (0 and 1) were replaced with NaN.
  - The data was grouped by TAXCLASS and BORO, and missing values were filled with the group mean (161,131 filled, 2 remaining).
  - For remaining missing values, the data was grouped by TAXCLASS alone, and the group mean was used for imputation (0 remaining).

- **Lot Depth (LTDEPTH)**
  - There were 161,715 missing values in the LTDEPTH.
  - The invalid values (0 and 1) were replaced with NaN.
  - Missing values were first filled using group means based on TAXCLASS and BORO (161,713 filled, 2 remaining).

- For unresolved missing values, grouping by TAXCLASS alone and imputing with the mean resolved all remaining values (0 remaining).

- **Building Width (BLDFRONT)**
  - There were 75 missing values in the BLDFRONT.
  - All missing values were successfully filled using group means based on TAXCLASS, BORO, and BLDGCL, reducing the missing count to 0.
  - As a precaution, additional methods were implemented using group means by TAXCLASS and BORO, followed by TAXCLASS alone, confirming that no missing values remained.

- **Building Depth (BLDDEPTH)**
  - There were 58 missing values in BLDDEPTH.
  - All missing values were successfully filled using group means based on TAXCLASS, BORO, and BLDGCL, reducing the missing count to 0.
  - As a precaution, additional methods were applied by filling with group means based on TAXCLASS and BORO, followed by TAXCLASS alone.

IV. Variable Creation

1. Purpose

The purpose of variable creation is to enhance anomaly detection by introducing derived metrics that identify discrepancies between property characteristics and valuations. These variables, including ratios, inverse measures, and regional averages, help pinpoint properties with unusual valuations for further investigation.

2. Variable Creation Summary Table

No.	Variable	Description	# Variables Created
1	**ltsize**	• Lot Size • Lot Width(LTFRONT) * Lot Depth(LTDEPTH)	1
2	**bldsize**	• Building Size • Building Width(BLDFRONT) * Building Depth(BLDDEPTH)	1
3	**bldvol**	• Building Volume • Building Size(bldsize) * Stories(STORIES)	1
4	**r1-3**	• Market Value Ratio • FULLVAL divided by ltsize(r1), bldsize(r2), bldvol(r3)	3
5	**r4-6**	• Actual Land Value Ratio • AVLAND divided by ltsize(r1), bldsize(r2), bldvol(r3)	3
6	**r7-9**	• Actual Total Value Ratio • AVTOT divided by ltsize(r1), bldsize(r2), bldvol(r3)	3
7	**r1-9 inverse**	• Inverse Ratios for r1–r9 to Detect Low Outliers • After Scaling, 1 / (r1–r9 + epsilon)	9
8	**Zip5_mean**	• Mean of r1–r9 for Each Zip Code	1
9	**taxclass_mean**	• Mean of r1–r9 for Each Tax Class	1
10	**value_ratio**	• Market Value to Assessment Ratio • Ratio of Market Value(FULLVAL) to the sum of Actual Land Value(AVLAND) and Actual Total Value(AVTOT)	1
11	**size_ratio**	• Ratio of Building Size(bldsize) to Lot Size(ltsize)	1

3. Variable Motivation

The anomalies or fraud cases are based on the premise that property valuation should reasonably align with the property size. To identify potential discrepancies, variables were created as ratios comparing physical dimensions of properties (e.g., lot size, building size, volume) to valuation and assessment metrics (e.g., full market value, land assessment value).

1) Variable No. 1 ~ 3

To calculate the physical dimensions of properties, three variables were created. ‘Lot size’ is calculated as the product of lot width (LTFRONT) and lot depth (LTDEPTH), representing the total area of the property. ‘Building size’ is calculated as the product of building width (BLDFRONT) and building depth (BLDDEPTH), representing the footprint of the building.

‘Building volume’ is calculated as the product of building size and the number of stories (STORIES), representing the total space within the building.

2) Variable No. 4 ~ 6
Ratios of full market value, actual land value, and actual total value to physical dimensions measure the value per unit of lot size, building size, or building volume. These ratios help identify properties with disproportionately high or low valuations relative to their physical characteristics, indicating potential anomalies or errors.

3) Variable No. 7
To identify outliers in the property valuation ratios, the following steps was applied to the 9 key ratio variables (r-r9):

- Scaling: Each variable was divided by its median value to standardize the variables and make them comparable.
- Inverses: For each variable, an inverse was calculated as 1/(variable + epsilon) to avoid division by zero. This transformation ensures that when a variable is close to zero (very low), its inverse becomes very large, converting low outliers into high outliers for easier detection.
- Consolidation: For each variable, the maximum value between the original variable and its inverse was retained. This ensures that the final variable captures extreme values in both directions. High outliers are naturally retained from the original variable. Low outliers are captured as large values in the inverse variable.
- Cleanup: The intermediate inverse columns were removed, leaving only the consolidated variables.

By the end of this process, the dataset retains the 9 adjusted variables (r1-r9), which highlight both very high and very low outliers in property valuation ratios. These outliers are likely candidates for further investigation as potential anomalies or fraudulent cases.

4) Variable No. 8 ~ 9
The mean of each property’s valuation ratios (r1-r9) was calculated to determine whether a property’s valuation ratios align with the typical values for its ZIP code area or tax class.

5) Variable No. 10 ~ 11
Value ratio measures the relationship between full market value (FULLVAL) and the sum of land and total assessments (AVLAND + AVTOT), highlighting valuation anomalies. Size ratio captures the ratio of building size to lot size, identifying properties with disproportionately small or large buildings.

V. Dimensionality Reduction

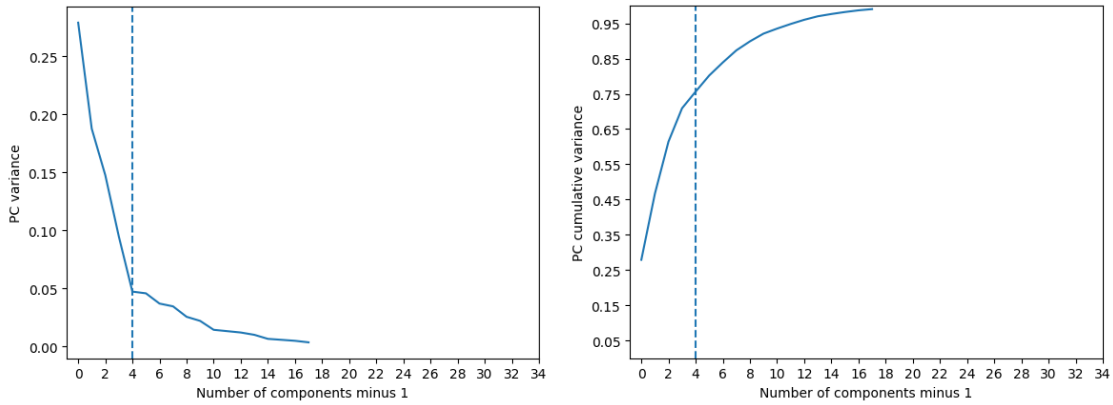
1. Purpose

To detect anomalies, it was necessary to handle high dimensionality and correlations among variables. This was achieved by reducing the data's complexity using Principal Component Analysis (PCA).

2. Process

1) Z-scaling Variables: Standardized all variables to have a mean of 0 and a standard deviation of 1 to prepare for PCA.

2) PCA and Scree Plot: Identified 5 principal components to retain 75% of the variance, balancing simplicity and information.



3) Redo PCA with Selected Components: Re-performed PCA with the top 5 components (those explaining the majority of variance) to reduce dimensionality while retaining key information.

4) Re-Z-scaling PCs: Standardized the selected PCs to ensure equal importance and improve distance-based anomaly detection.

VI. Anomaly Detection Algorithms

1. Purpose

Two unsupervised fraud detection algorithms were used to generate fraud scores (Score 1 and Score 2). These methods leveraged the reduced-dimensional PCA-transformed data.

2. Process

1) Z-Score Outliers

- Calculating the unsupervised fraud score (Score1)
    - The Minkowski Distance (p=2, Euclidean) was used to calculate the distance of each observation from the origin in PCA-transformed space.
    - Larger distances indicated significant deviations from the norm, producing Score 1.

2) Autoencoder

- Create and Train the Autoencoder
    - A simple neural network was built using MLPRegressor to act as the autoencoder.
    - It compresses the PCA-transformed data into a smaller representation and learns to reconstruct it.

- Predict and Calculate Reconstruction Error
    - The autoencoder was used to recreate the input data. The reconstruction error (difference between original and reconstructed data) was calculated.
    - The Minkowski Distance (p=2, Euclidean) was used to measure this error, producing Score 2.

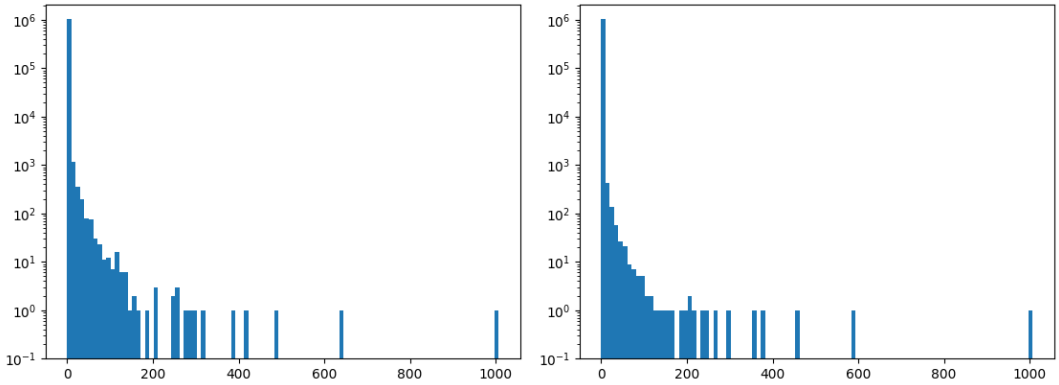
VII. Results

To detect anomalies, two unsupervised fraud detection algorithms were employed, generating two fraud scores: Score 1 and Score 2.

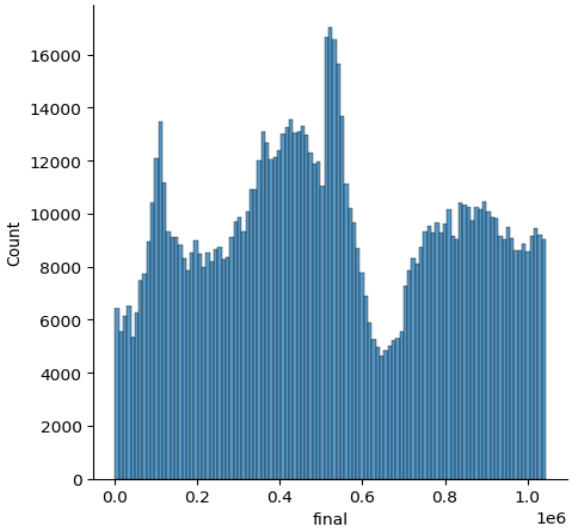
1. Score Integration

The calculated fraud scores were ranked to simplify comparisons, with higher ranks indicating more unusual records. A final fraud score was obtained by averaging the ranks of Score 1 and Score 2.
The distributions of the two scores are displayed in the graphs above, with flatter plots indicating greater similarity between the two scores. This similarity suggests consistent anomaly detection across the two algorithms.

The graphs show the distribution of Score 1 (left) and Score 2 (right).

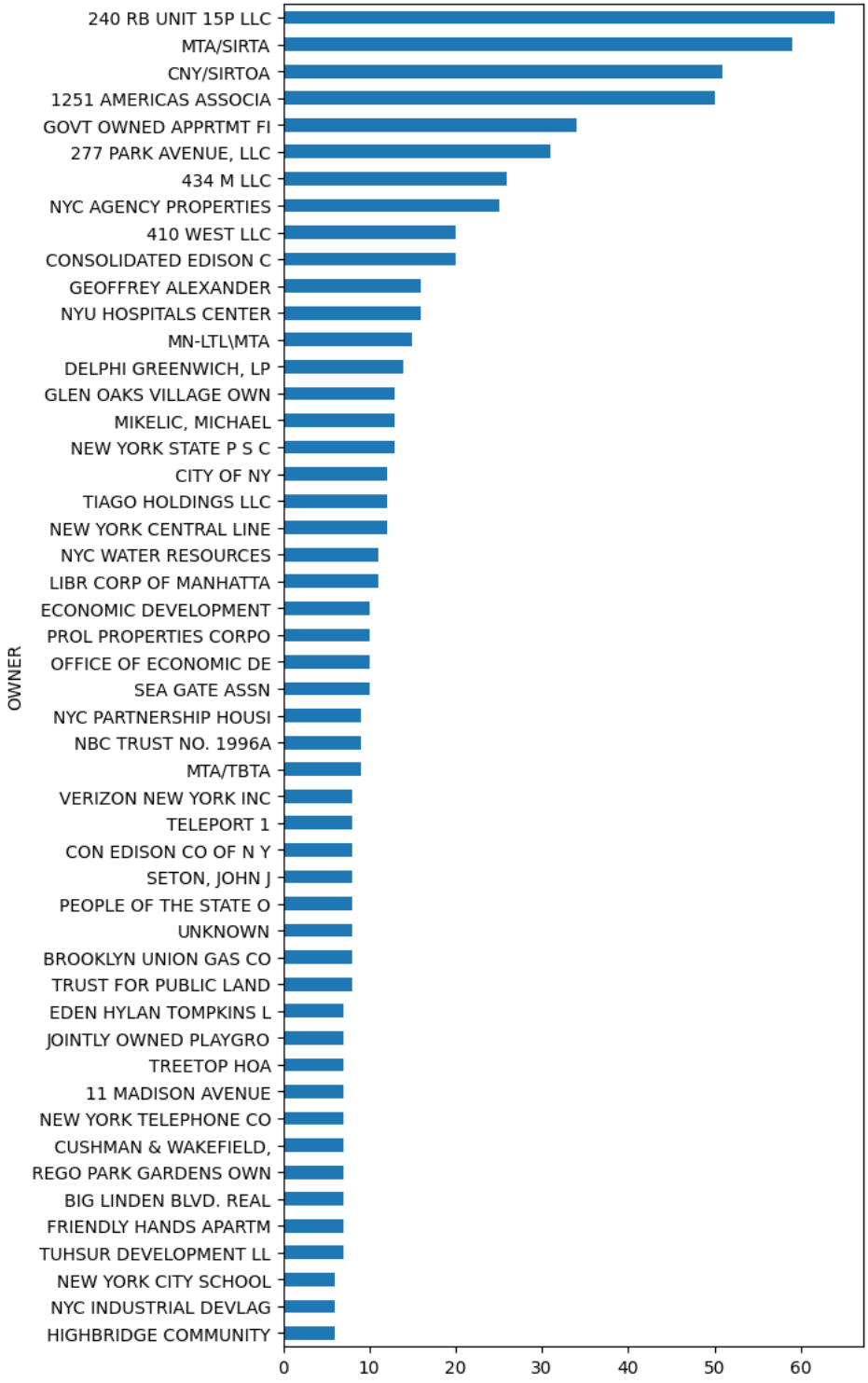


The flatter the plot, the more similar the two scores are, indicating they are relatively similar.

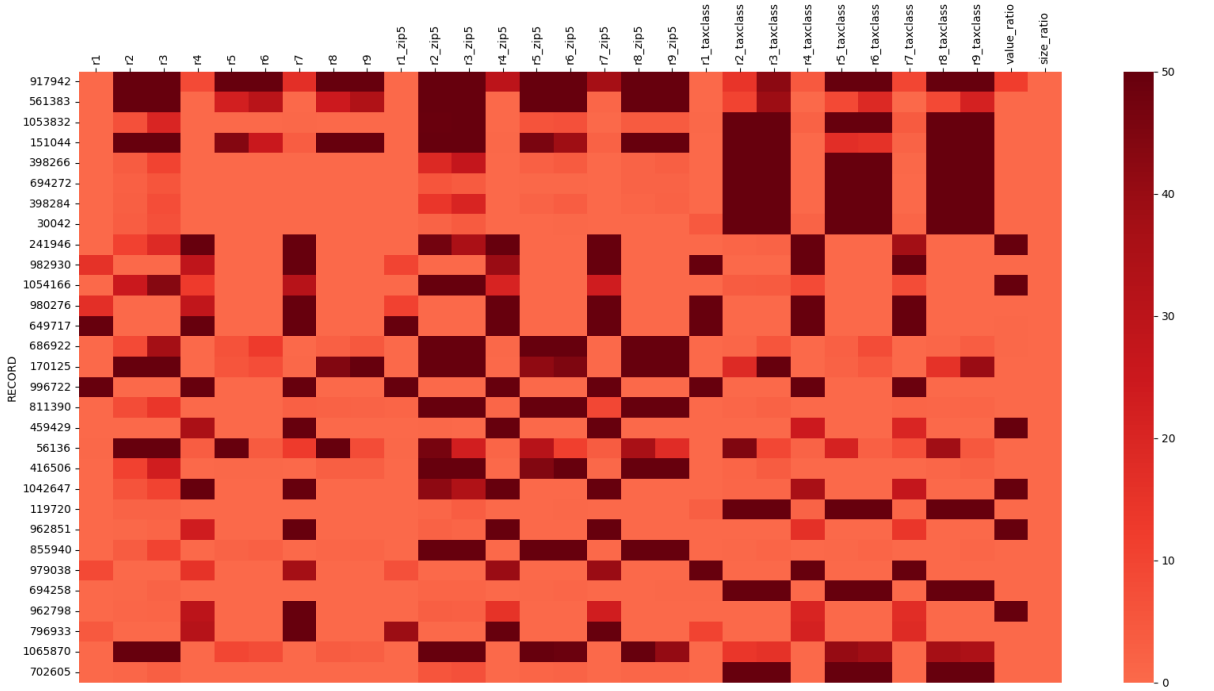


2. Identifying and Visualizing Top Anomalies

The highest-ranking records based on the final fraud scores were selected and integrated back into the original dataset for further investigation. A horizontal bar chart was created to highlight the top 50 owners associated with the most anomalous properties.



A heatmap of the top 30 anomalies was created, showing the absolute values of all key variables. Dark red cells represent high absolute values, indicating significant deviations or anomalies. For dark rows, the records with multiple dark cells across variables are more likely to be anomalies. As for dark columns, variables with frequent dark cells across records are the most impactful in identifying anomalies.



3. Case Studies

1) LOGAN PROPERTY, INC. (917942)

The property shows unusual patterns in r5 (1000), r6 (1019), r8 (984), and r9 (1014), particularly regarding building size and volume. According to the original data, the building size is problematic, as both building width (BLDFRONT) and building depth (BLDDEPTH) are recorded as 0. Consequently, the calculated building volume is 0, despite the property having 3 stories (STORIES). This discrepancy indicates an anomaly in the reported building metrics.

Now, the market value (FULLVAL) is set at $374,019,883, while the actual land value (AVLAND) is $1,792,808,947 and the actual total value (AVTOT) is $4,668,308,947. Logically, the market value should represent the total assessed value of the property based on its market worth, meaning it should be greater than or equal to both the actual land value and total value. However, in this case, the market value is significantly lower than both AVLAND and AVTOT, which is a clear inconsistency. This discrepancy raises the possibility that LOGAN PROPERTY, INC. may have intentionally underreported the market value to reduce tax obligations.

2) YILDIZ HOLDING A.S. (561383)

The irregularities in this case are especially highlighted when viewed in the context of the property's ZIP5 region. Variables r2 (589), r3 (654), r5 (494), r6 (555), r8 (557), and r9 (585), all related to building size and volume, show significant deviations from typical values observed in this ZIP5 area. Notably, both the building's front (BLDFRONT) and depth (BLDDEPTH) are recorded as 0, despite the property being listed with 2 stories (STORIES), resulting in a calculated building size and volume of 0. Additionally, there is a clear misalignment between the market value and the assessed values. The market value is listed as $258,000,000, while the actual land value (AVLAND) is $40,590,000, and the actual total value (AVTOT) is $116,100,000. Given the unusual patterns within the ZIP5 region for these variables, the property stands out as an outlier.

3) MARKOW, REGINA (1053832)

This property exhibits a similar trend of anomalies, particularly when analyzed within its tax class. Variables r2, r3, r5, r6, r8, and r9 show irregularities ranging from 188 to 666 within Tax Class 1. Adding to the irregularities, both the building width (BLDFRONT) and depth (BLDDEPTH) are recorded as 0, while the property is listed as having 2 stories (STORIES), resulting in a calculated building size and volume of 0. The market value is reported as $20,300,000, while the actual land value (AVLAND) is $163,536, and the actual total value (AVTOT) is $172,009. This extreme discrepancy is even more pronounced compared to the previous cases, raising significant doubts about the accuracy of the reported values and warranting closer scrutiny of this property.

VIII. Summary

This project focused on detecting property tax fraud in New York City using an unsupervised approach to analyze approximately 1 million property records. A systematic pipeline was implemented, beginning with data cleaning to ensure a high-quality dataset. Missing values were imputed, outliers were preserved as potential indicators of fraud, and government-owned properties were excluded to focus on private property transactions. Through variable creation, innovative features like valuation ratios, inverse metrics, and regional averages were developed to capture anomalies in property valuation relative to physical and locational characteristics.

Dimensionality reduction via Principal Component Analysis (PCA) was employed to address the challenges of high dimensionality and correlations among variables, condensing the data into five principal components while retaining 75% of the variance. Two unsupervised anomaly detection methods were then applied: Z-Score Outliers, which used Minkowski distance in PCA-transformed space, and an Autoencoder, which calculated reconstruction errors to highlight unusual records. Fraud scores from both methods were ranked, combined, and visualized to identify the most anomalous properties and influential variables.

The results effectively pinpointed records with significant deviations, providing actionable insights into potential fraud. Visualizations such as bar charts and heatmaps facilitated further investigation, highlighting patterns among anomalous properties and their owners. For instance, rows with consistently dark cells across multiple variables in heatmaps indicated records worthy of deeper examination, while columns with frequent anomalies pinpointed impactful variables.

Expert feedback is essential to refining the fraud model by defining what constitutes "unusual" behavior in the context of property tax fraud. Through interviews with domain experts, potential fraud scenarios and dynamics can be identified, guiding the creation of variables that capture these insights, such as valuation ratios or transaction anomalies. These variables can be transformed and combined to highlight potential fraud indicators while reducing redundancy and correlations through PCA. By scaling the reduced dimensions effectively, the model focuses on meaningful patterns, enabling anomaly detection through techniques like Z-Score Outliers or Autoencoder-based reconstruction errors. This iterative approach ensures the model evolves to target fraud-relevant anomalies, improving detection accuracy and practical application.